

Орися ДЕМСЬКА-КУЛЬЧИЦЬКА,
канд. філол. наук

СИСТЕМА КОДУВАННЯ ПЕРВИННИХ ДАНИХ КОРПУСУ

Застосування комп'ютера у лінгвістиці не лише модифікувало систему дослідних підходів до мови, а й призвело до формування нових методик оформлення мовного матеріалу для наступного лінгвістичного опрацювання -- вивчення, аналізу, опису, систематизації etc. Власне як нову модель організації мовного матеріалу розглядаємо корпус текстів *машиночитане, стандартно організоване зібрання репрезентативних для певної мови, діалекту або іншої підмножин(и) мов(и) писемних або усних текстів, призначених для лінгвістичного аналізу й опису, відібраних і впорядкованих згідно з експліцитними екстра- та інтралінгвальними критеріями.*

Важливою диференційною ознакою текстового зібрання іменованого корпусом стосовно інших електронних об'єктів, сформованих із текстів природної мови, є його стандартна тричленна структура, яка передбачає обов'язкову наявність трьох технологічних елементів, а саме: (1) електронного заголовка, (2) визначення типу документа, більше відомого як DTD і (3) первинних даних, якими є закодовані згідно з вимогами стандартів кодування корпусу тексти довільної мови.

Базовими стандартами, які забезпечують кодування первинних даних у корпусі, вважаються: SGML (Standard Generalized Markup Language) як першостандарт, а також документи, збудовані на принципі застосування концепції мовної незалежності та описової розмітки (SGML, - TEI (Text Encoding Initiative) і CES (Corpus Encoding Standard)?, створені спеціально для кодування корпусних ресурсів незалежно від мови текстів, які входять до корпусу.

Вибираючи між форматами кодування первинних даних TEI та CES, схиляємося до принципів TEI, оскільки цей стандарт забезпечує оптимальну збалансованість між загальною моделлю подання природної мови і нескладною реалізацією кодування. Також TEI оперує великим набором засобів для подання як лінгвальної, так і металінгвальної інформації.

Передумовою розроблення системи ТЕІ стало існування великої кількості несумісних систем кодування і розширення сфери застосування електронних текстів. Базовими принципами системи визначено: а) можливість досягати у тексті ефектів, необхідних для наукових досліджень різного типу; б) простота, чіткість і конкретність; в) нескладність для використання без спеціалізованого програмного забезпечення; г) можливість точного визначення та ефективного програмного оброблення текстів; ґ) можливість розширень, визначених користувачем; д) узгодженість з чинними і новостворюваними стандартами.

Згідно з Принципами ТЕІ кодування первинних даних у корпусі передбачає описове подання: (1) структури електронного тексту: частин, заголовків, абзаців, цитат, речень; (2) незалежно від мови типових текстових елементів, які можуть мати як інтратекстовий, так і екстратекстовий характер; (3) типових одиниць лінгвального рівня, які виділено у тексті написанням, наприклад: оніми, записані з великої літери, виділені курсивом терміни тощо.

Кодування глобальної структури первинних даних (маркер <text>) відбувається за допомогою елементів <front>, <group>, <body> і <back>, які означають:

о <front> - довільна вступна інформація, розміщена перед основним текстом і йдеться про заголовки, титульний лист, передмови, присвяти тощо; о <group> - кілька згрупованих монотекстів; о <body> - основна частина моно- чи політексту, крім тексто-вої інформації вступної або кінцевої частин; о <back> - довільна текстова інформація, розміщена після основного тексту: додатки, дати, підписи, PS, тощо.

Перший і останній з елементів є факультативним. Закодований політекст повинен відповідати моделі:

```
<TEI2>
<teiHeader> [ інформація заголовка об'єднаного тексту ] </teiHeader>
<text>
  <front> [ вступ об'єднаного тексту ] </front>
  <group>
    <text>
      <front> [ вступ до першого тексту ] </front>
      <body> [ тіло першого тексту ] </body>
      <back> [ закінчення першого тексту ] </back>
    </text>
    <text>
      <front> [ вступ до другого тексту ] </front>
      <body> [ тіло другого тексту ] </body>
      <back> [ закінчення другого тексту ] </back>
    </text>
  </group>
</text>
```

```
[ інші тексти або групи текстів ] </group>
<back> [ закінчення об'єднаного тексту ] </back> </text>
</TEI2>
```

Після кодування глобальної структури первинних даних, слід обов'язково задати формальні ознаки тексту, важливі для наступної лінгвістичної інтерпретації. До таких ознак належать зокрема специфіка редакторської розмітки, яка зазвичай полягає у виділенні певних елементів тексту, наприклад, заголовка і підзаголовків, цитат, іншомовних слів, термінів тощо.

Візуально виділення може бути передано різними способами, наприклад, погрубленням і/або курсивом малими літерами, погрубленням і/або курсивом великими літерами, петітом, розрядкою літер тощо. Застосовуючи Принцип ТЕІ, для неінтерпретованого типографського виділення використаємо елемент <hi>, який маркує слово або фразу, що графічно відрізняється від основного тексту і причина цього виділення невідома або не ідентифікована. Елемент <hi> оперує атрибутом *rend* для вказівки на тип виділення текстового фрагменту. Наприклад, наведений нижче текст з виділеними словами 'лексична', 'картотека', 'художні', 'твори', 'діалектний', 'матеріал', 'ЛІК', 'єдиним' слід кодувати:

Лексична картотека - це зібрання карток-ілюстрацій лексико-фразеологічних, стилістичних, діалектних та ономастичних багатств української мови, зібраних на основі текстових матеріалів (художні твори, діалектний матеріал тощо) української мови XIX-XX ст. ЛІК Інституту української мови НАН України є єдиним у світі такого типу зібранням і складається з ...

```
<text>
  <p>
    <s><hi rend=bold>Лексична</hi> <hi rend=bold>картотека</hi> - це зібрання
    карток-ілюстрацій лексико-фразеологічних, стилістичних, діалектних та ономастичних багатств
    української мови, зібраних на основі текстових матеріалів (<hi rend=italic>художні</hi>, <hi
    rend=italic>твори</hi>, <hi rend=italic>діалектний</hi> <hi rend=italic>матеріал</hi> тощо)
    української мови XIX-XX ст. </s> <s><hi rend=italic>bold>ЛІК</hi> Інституту української мови НАН
    України є <hi rend=italic>rend= bold>єдиним</hi> у світі такого типу зібранням і складається з </s>
  </p>
</text>
```

Якщо ж причина виділення текстового фрагменту з'ясована або відома, глобальний елемент <hi> замінюється на спеціалізовані <emph>, <foreign>, <mentioned>, <term> і <title>, які означають: о <emph> - виділення емпізи; о <foreign> - виділення запозичення;

о <mentioned> - виділення цитованого або ілюстративного матеріалу;
 о <term> - виділення терміна;
 о <title> - виділення назви / заголовка / підзаголовка, типологія яких експлікується через атрибути *level* і *type* з відповідними значеннями: *level* - зазначає тип заголовка (назва статті, книги, журналу, серії або неопублікованого матеріалу); *type* - класифікує назви відповідно до прийнятої типології через: ABBREVIATED - аббревіація, MAIN - основна назва, SUBORDINATE - підзаголовок і назва частин та PARALLEL - альтернативні назви. Наприклад, кодування виділення терміна-запозичення в наведеному нижче уривку матиме такий вигляд:

Документ є міжнародним стандартом на опис розміченого електронного тексту. Точніше, SGML - це метамова, тобто, засіб формального опису мови...

```
<p>
  <s>Документ є міжнародним стандартом на опис
  розміченого електронного тексту.</s> <s>Точніше,
  SGML&mdash;це <term type=bold>метамова</term>
  (<foreignlang='en'>metalanguage</f oreign>),
  тобто, засіб формального опису мови</s>
</p>
```

Процес кодування первинних даних має багато спільного з процесом традиційного редагування. Редакторські виправлення у процесі кодування первинних даних повинні перш за все охоплювати: а) описки і помилки переписувачів у давніх текстах, б) сучасні описки; в) помилкове дублювання одного і того ж слова у тексті як історичному, так і сучасному, г) граматичні русизми, г) суржикізми, д) орфографічні помилки, е) семантичні ляпи.

Кодування редакторських змін доцільно реалізувати за допомогою таких елементів, детермінованих у TEI, як <corr>, <sic>, <orig> і <reg>, що означають:

о <corr> - засвідчує правильний запис, який в оригіналі наведений з явними помилками.
 о <sic> - містить текст, який доцільно відтворити без змін, незважаючи на його явну некоректність, помилковість чи неточність.
 о <orig> - запис, зафіксований в оригіналі, можливо навіть помилковий;
 о <reg> - виправлений запис з атрибутами *orig* - не виправлений варіант тексту-джерела і *resp* - особа, відповідальна виправлення.

Наприклад, уривок з твору Лесі Українки „На полі крові”:

Дідок-прочанин іде поз нивку стежкою, що звертає в бік з великого Єрусалимського шляху ...

Вимагає таких коментарів: у цьому тексті, по-перше, орфографічна помилка допущена у слові *поз*. яке правильно повинно писатися *повз*, і, по-друге, допущена помилка у написанні прислівника *місця вбік*, який проаналізовано як іменник *бік* з прийменником *в* і записано окремо. Щоби коректно закодувати цей текст слід зробити таке:

```
<text>
  <p>
    <s>Дідок-прочанин іде <reg orig>повз</reg> нивку стежкою, що звертає <reg orig>в
    бік>вбік</reg> з великого Єрусалимського шляху ... </s>
  </p>
</text>
```

У процесі кодування первинних даних може виникнути необхідність усувати ще такі помилки, як, наприклад, механічні пропуски у тексті, повтори, нерозбірливий запис або затерті чи знищені частини тексту тощо. Для цього в TEI передбачено використовувати елементи <add>, <gap>, і <unclear> з відповідною семантикою: о <add> - текстова одиниця (буква, слово, фраза), узуповнена на місці пропуску; о <gap> - місце пропуску; о - містить текст вилученого матеріалу; о <unclear> - текстова одиниця, яка не піддається ідентифікації через технічні ушкодження тексту. Усі ці елементи оперують певними атрибутами, за допомогою яких експлікують типологію помилок і принципи їх виправлення під час кодування первинних даних. Наприклад, якщо оригінал електронного тексту:

Тонкі ніжні ніжні берези поперепліталися з поважними дубами дубами і ясними літніми ночами блистять, мов у срібло одягнені.

то його кодування передбачатиме усунення помилкового повтору:

```
<text>
  <p>
    <s>Тонкі ніжні <del hand=LB>ніжні</del> берези
    поперепліталися з поважними дубами <del
    hand=LB>дубами</del> і ясними літніми ночами блистять,
    мов у срібло одягнені.</s>
  </p>
</text>
```

Усі попередньо розглянуті одиниці кодування пов'язані із загальною структурою полотна тексту, натомість наступна група одиниць, які підлягають кодуванню на рівні первинних даних, згідно з Принципами TEI зосереджені в межах абзацу. Структурними елементами первинних даних у межах текстового абзацу зазвичай вважаються цитати, списки, таблиці, графічні зображення, формули, адреси, віршові рядки, речення. І залежно від аплікативних вимог корпусу, прийнято детермінувати набір елементів текстового абзацу, які оброблятимуться програмно, для кожного конкретного корпусу чи підкорпусу. Так, у корпусі української мови на рівні абзацу в первинних даних пропонуємо кодувати (1) цитати, (2) віршові рядки, (3) списки, (4) таблиці, (5) адреси і (6) речення.

Стандартно текстовий абзац кодується елементом <p>. У межах абзацу для кодування детермінованих вище одиниць призначені маркери <q>, <l>, <lg>, <sp>, <speaker>, <stage>, <list>, <table>, <address> і <s>, які означають:

о <q> (інтратекстова цитата) ідентифікує мову автора в межах власного твору, подану як цитату, наприклад:

На дні тенденції стосовно гуманітарних наук вказує І. Штерн: «**передовсім слід звернути увагу на різноманітність сучасних способів формування предметних галузей в гуманітарній галузі й цю вражаючу легкість, з якою вони виникають**» і далі зауважує, що особливе місце в гуманістиці займають, так звані, гібридні дисципліни.

<p>

<s>На дві тенденції стосовно гуманітарних наук вказує <name>І. Штерн</name>: <quote>передовсім слід звернути увагу на різноманітність сучасних способів формування предметних галузей в гуманітарній галузі й цю вражаючу легкість, з якою вони виникають </quote> і далі зауважує, що особливе місце в гуманістиці займають, так звані, гібридні дисципліни. </s>

</p>

Крім маркера <q>, залежно від типології цитати, можуть також використовуватися: <quote> - (інтертекстова цитата) ідентифікує текстовий матеріал, взятий з іншого тексту, оформлений цитатою; <cit> - цитата з бібліографічним посиланням; <mentioned> - фактичний матеріал, поданий у тексті як приклади; <soCalled> - фрагмент тексту, поданий після таких конструкцій як: „так би мовити“, „за словами Пана Х“, тощо.

У Принципах TEI передбачено кодування віршового та драматичного текстів і запропоновано це реалізовувати через <l>, <lg>, <sp>, <speaker>, <stage>, де:

о <l> - подає рядок віршового тексту, а метричну завершеність і незавершеність конкретного рядка специфікує атрибут *part*;

о <lg> - кодує групу віршових рядків, які формально становлять цілісну одиницю;

о <sp> - ідентифікує індивідуальне мовлення, організоване як вірш, в межах прозового тексту, з вказівкою на мовця через атрибут *who*;

о <speaker> - елемент, призначений для забезпечення інформації про власну назву мовця / мовців у драматичному тексті; о <stage> - довільна сценічна ремарка у драматичному тексті, тип якої ідентифікує атрибут *type*. У довільному корпусі очевидно з'явиться фактичний матеріал, який може мати форму списків або таблиць, наприклад, козацькі реєстри. TEI трактує, зокрема, список як впорядковану / неупорядковану послідовність текстових одиниць або глосарій. Схема кодування списку оперує набором елементів <hst>, <item>, <label>, <head>, <headLabel>, <headItem>, які означають:

о <list> - довільна послідовність одиниць, що складають список.

о <item> - структурний елемент списку;

о <label> - мітка, зв'язана зі структурним елементом списку;

в глосаріях маркує тлумачений термін;

о <head> - довільний заголовок або підпис списку чи його частини;

о <headLabel> - підпис мітки або терміна-мітки в глосарії чи структурному елементі списку;

о <headItem> - підпис структурних елементів списку або глосарію, або просто структурованого списку. Елемент <list> можна використовувати для маркування будь-якого списку: нумерованого, літерованого, символізованого або неміченого взагалі. Залежно від вимог оброблення списку, нумерація в ньому може бути а) пропущена, б) визначена за допомогою атрибута *n*, або в) розмічена тегом <label> як вміст елемента <list>. Наприклад:

```
<list>
<head>Короткий список</head>
<item>Перша позиція у списку.</item>
<item>Друга позиція у списку.</item>
<item>Третя позиція у списку.</item>
</list>
```

```
<list>
<head>Короткий список</head>
<item n=1>Перша позиція у списку.</item>
<item n=2>Друга позиція у списку.</item>
<item n=3>Третя позиція у списку.</item>
</list>
```

```
<list>
<head> Короткий список </head>
<label>1</label><item>Перша позиція у списку.</item>
<label>2</label><item>Друга позиція у списку.</item>
<label>3</label><item>Третя позиція у списку.</item>
</list>
```

Останнім з елементів абзацного рівня розглянемо речення. У схемі кодування TEI на рівні абзацу можливе, але не обов'язкове виділення речення, проте мета створення загальномовного корпусу, його призначення та лінгвістична орієнтованість ставить вимогу обов'язкового кодування цього елемента. Маркування орфографічного речення передбачено реалізувати шляхом застосування маркера `<s>` (речення) з можливими глобальними та індивідуальними атрибутами *type* - деталізує тип синтаксичного сегмента і *function* - деталізує специфіку функціонування синтаксичного сегмента, наприклад:

```
<p>
<s>Суверенітет <name type=place> України </name> поширюється на вся її територію.</s> <s>name
type=place>Україна</name> є унітарною державою.</s>
</p>
<p>
<s>Територія <name type=place>України</name> в межах існуючого кордону є цілісною і
недоторканою.</s>
</p>
```

У межах речення на рівні кодування первинних даних слід пам'ятати про маркування слів, записаних з великої літери, аббревіатур, чисел і розділових знаків.

Вживання великої і малої літери, наприклад, згідно з принципами українського правопису торкається опозиції „онім - апелятив". У первинних даних в межах групи слів-онімів виділяють усі типи власних назв і маркують їх через `<name>` (довільна власна назва) з атрибутом *type*, призначення якого типологізувати кодований онім, наприклад:

Валерій Шевчук розпочинає свій твір „Три листки за вікном" цитатою, з **Г.Сковороди**: «Світ неситий, коли не задовольняє. Вічність несити, коли не завдає жалю... А я, як був, так і тепер – подорожній!..».

```
<p>
<s><name type=prope, reg=Шевчук Валерій>Валерій Шевчук</name> розпочинає
свій твір <q>Три листки за вікном</q> цитатою з </name type=prope,
reg=Сковороди>Григорія</name> <quote> Світ неситий, коли не
задовольняє. Вічність несити, коли не завдає жало...
А я, як був, так і тепер – подорожній!..</quote>.</s>
```

Абревіатуру чи складноскорочене слово, не залежно від типологічних характеристик - ініціальні, звукові, складові чи слова-словосполучення, - за схемою TEI передбачено кодувати короткий та повний запис абревіатури в межах елемента `<abbr>`, який маркує довільне скорочення з

формальною семантизацією довільної аббревіатури через атрибут *type* - детермінує тип скорочення відповідно до прийнятої TEI класифікації із значеннями: CONTRACTION - стягнена форма, SUSPENSION - пропуск, три крапки, SUPERScription - верхній індекс і ACRONYM - акронім; можливі також значення TITLE - назва в адресі, GEOGRAPHIC - географічна назва, ORGANIZATION - назва організації тощо.

Наприклад, фрагмент тексту '*США - це супердержава сучасного світу*' можна закодувати так:

```
<p>
<s><abbr type=gaogr>США</abbr> це супердержава сучасного світу.</s>
</p>
```

На рівні речення у первинних даних ще одним важливим об'єктом кодування є число, яке може бути записане за допомогою буквених або цифрових символів. Не залежно від принципу фіксації числової інформації, її кодування можна здійснювати за допомогою елементів `<num>`, `<date>` і `<time>`, де:

о `<num>` - довільно записане число.

Атрибути: *type*, який експлікує тип числового значення через: FRACTION - дріб, ORDINAL - порядковий числівник, PERCENTAGE - відсоток і CARDINAL - абсолютне число, і *value* - стандартне подання значення числа.

о `<date>` - дата у довільному форматі запису.

Атрибути цього елемента *calendar*-визначає систему числення чи календар, якому відповідає дата, і *value* - стандартно подає значення лати, звичайно у форматі „рік - місяць - день"-;

о `<time>` - часова інформація в межах доби, подана у довільному форматі запису.

Зуважимо, що атрибут *value* в межах елементів `<date>` і `<time>` визначає стандартний запис через ISO 8601.

Наприклад:

```
... <num type=cardinal value='21'>двадцять один</num>...
... <num type=percentage value='10'>десять</num>...
... <num type=percentage value='10'>10%</num>...
... <num type=ordinal value='5'>5-ий</num>...
... <date value='1980-02-21'>21 люте 1980</date>...
... <date value='1990'>1990</date>...
... <date value='1990-09'>Вересень 1990</date>
```

Кодування пунктуаційних символів у корпусі може мати факультативний характер. Винятком повинні бути дані для синтаксичних дос-

ліджень. Зауважимо, що семантика розділових знаків залежно від позиції і відступів перед і після знака є різною. Загалом пунктуаційні знаки прийнято кодувати такими символами (див. табл. 1).

Табл. 1: *Набір кодів для кодування пунктуаційних знаків*

Символ	Значення символу	Знак
<FSTOP>	крапка - full-stop	.
<PER>	три крапки - period	...
<COM>	кома - comma	,
<SC>	крапка з комою - semi-colon	;
<COL>	двокрапка - colon	:
<QUEST>	знак питання - question mark	?
<EXCL>	знак оклику - exclamation mark	!
<LPAR>	ліва (початкова) кругла дужка - left (opening) parenthesis	(
<RPAR>	права (кінцева) кругла дужка - right (closing) parenthesis)
<ODBLQ>	подвійні початкові лапки - open double-quotes	«; „
<CDBLQ>	подвійні кінцеві лапки – close double-quotes	»; "
<&MDASH>	тире - dash	—
<HYPH>	дефіс - hyphen	-
<SLASH>	скісна риска - slash	/
<LSQR>	ліва (початкова) квадратна дужка - left (opening) square bracket	[
<RSQR>	права (кінцева) квадратна дужка - right (closing) square bracket]
<SYM>	символи	~; *

Отже, навіть доволі фрагментарний огляд системи кодування первинних даних, запропонований Принципами TEI, засвідчує, що довільний текст природної мови, закодований стандартно, перетворюється на стандартний об'єкт IT-середовища з можливістю багатократного та різноаспектного застосування не лише у лінгвістичних студіях. Крім того, підходи TEI до кодування первинних даних корпусу забезпечують уніфіковане і стандартне оброблення тексту шляхом застосування описової розмітки таких одиниць тексту як заголовки, абзац, список, слово тощо; збереження зв'язку між оригінальними і закодованими даними; нескладну реалізацію кодування, причому без спеціалізованого програмного забезпечення.

ПРИМІТКИ

1. Поза увагою залишаємо найостанніші версії цих документів.

ЛІТЕРАТУРА

1. *Sperberg-McQueen C M., Burnard L.* Guidelines for Electronic Text Encoding and Interchange. - <http://www.hcu.ox.ac.uk/TEI/P4X/index.html>, 2001.
2. *Ide N.* Corpus Encoding Standard. - <http://lpl.univ.-aix.fr/projects/multext/CES>, 2000.

АНОТАЦІЯ

У статті розглянуті проблеми кодування текстових даних, організованих як корпус текстів природної мови. Головним чином мова йде про принципи розмітки глобальної структури первинних даних, а також специфіка корпусного тексту.

Ключові слова: *корпус, первинні дані, кодування корпусу, Принципи TEI.*

SUMMARY

The article deals with the problems of the texts encoding, which is organizing as a natural language corpus. Manly we discussed the marking principles of the global structure of primary data and others corpus texts specificities.

Key words: *corpus, primary data, corpus encoding, TEI Gaudiness.*