ЗАСТОСУВАННЯ ПРИНЦИПІВ TEI ДО КОДУВАННЯ ТЕКСТОВИХ КОРПУСНИХ ДАНИХ

Орися Демська-Кульчицька

Інституту української мови НАН України,
01001 Київ, вул. Грушевського, 4,
факс (044) 229-56-19; тел. (044) 229-60-17
E-mail: odk@ukr.net

Рассмотрены проблемы кодирования текстовых данных, организованных как корпус текстов естественного языка, в частности украинского. Главным образом речь идет о принципах разметки глобальной структуры первичных данных, а также специфик корпусного текста.

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

Один із методів подання лінгвального ресурсу в IT-середовищі є корпусний метод, який передбачає організацію реального мовного матеріалу з наступним його кодуванням на рівні структури тексту і граматичної, лексичної, дискурсної тощо семантики конститутивних одиниць тексту, тобто побудови корпусу текстів природної мови як певної лінгвістичної моделі організації природномовного матеріалу для лінгвістичних досліджень і програмного застосування.

Входження природної мови у сферу інформаційних технологій ставить вимогу як лінгвістичного, так і особливо технологічного її забезпечення. Лінгвістичне і технологічне взаємодетермінуються. Некоректні рішення на рівні лінгвістичному призводять до некоректних реалізацій на рівні технологічному і навпаки. Тому стоїть завдання створити умови, за яких можливо було би послідовно, стабільно, однозначно і коректно обробляти природну мову як обов'язковий елемент сучасних інформаційних технологій. Одне з рішень лежить у сфері стандартності, яка власне створює передумови цього послідовного, стабільного, однозначного, узаконеного оброблення даних природної мови, не залежно від самої мови. Стандартність подання природної мови у комп'ютерному середовищі, крім вимог програмного характеру, зумовлена також необхідністю багатократного використання мовних ресурсів як емпіричної бази лінгвістичного дослідження, динамікою еволюційних процесів у техніці та програмних ресурсах оброблення даних природної мови і полілінгвальністю світу. Відповідно принцип стандартності мотивує функціонування стандартів, призначення яких забезпечити поліаплікабельність лінгвальних ресурсів. Базовим, вихідним стандартом, першостандартом тут є SGML (Standard Generalized Markup Language) як першостандарт та документи, які збудовані на принципі застосування концепції 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 [1, 2].

Вибираючи між форматом TEI та CES кодування первинних даних, схиляємося до принципів TEI, оскільки цей стандарт забезпечує оптимальну збалансованість між загальною моделлю подання природної мови і нескладною реалізацією кодування. Крім того, що TEI оперує великим набором засобів для подання як лінгвальної, так і металінгвальної інформації, його застосовують у найбільших проектах побудови корпусів національного типу, що дозволяє використовувати уже наявні засоби, призначені для обробки корпусних даних, поданих у цьому форматі.

Загалом TEI як ініціатива кодування тексту є міжнародним і міждисциплінарним стандартом подання усіх типів текстів, функціональних у бібліотечній, музейній, видавничій справах, мовознавстві, шляхом використання максимально виразної і мінімально застарілої схеми кодування. Сумісність прийнятої схеми кодування корпусу і загальної системи TEI визначають принципи кодування і обміну електронними текстами (Guidelines for Electronic Text Encoding and Interchange). Схема кодування TEI використовує стандартну мову узагальненої розмітки, тобто SGML. Це хоча й ускладнює саму систему, проте робить її універсальною, оскільки уникає принципових відмінностей між різними схемами розмітки текстів природної мови і таким чином уможливлює оброблення TEI-сумісних текстів будь-яким програмним забезпеченням загального призначення, що працює з SGML. Принципи TEI спеціально розроблені для різних застосувань та дисциплін і тому можуть обробляти не лише великі текстові одиниці, але й бути максимально узагальнювальними і гнучкими. Принципове призначення TEI – забезпечити коректний обмін текстової інформації, яка має електронний вигляд. Крім того, TEI забезпечує засоби експлікації архітектоніки тексту з метою спрощення його оброблення програмними засобами, які працюють на різних комп'ютерних платформах і у різних технологічних середовищах. Передумовою розробки системи TEI 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.

Уперше принципи системи TEI опубліковано у 1994 році. Впродовж 1990–92 років з'явився ряд драфтів про принципи TEI: 1990 (P 1), 1990 (P 1.1), 1992 (P 2) 1990, а з 1994 по 2002 рік їх опубліковано як стандарти кодування і обміну електронними текстами (Guidelines for Electronic Text Encoding and Interchange: 1994 (P3), 1999 (P3.1), 2002 (P4)), які складаються з (а) синтаксису: набір SGML/XML DTD і (б) семантики: документація TEI обумовлює розмітку структури тексту через набір тегів і TEI електронний заголовок. У 2001 році TEI стає консорціумом, об'єднуючи сили дослідників у галузі комп'ютерного оброблення природних мов. Загалом на сьогодні 88 проектів реалізовано або реалізовуються, використовуючи принципи TEI.

Крім повного варіанту принципів TEI, який визначає кілька сотень елементів SGML, доволі поширеним є документ TEI Lite, призначення якого забезпечити так званий стартовий набір елементів SGML для користувацького кодування тексту. TEI Lite охоплює ядро набору тегів TEI, обробляє достатньо великий діапазон текстів, уможливлює створення нових документів для кодування текстів, його можуть обробляти програмні засоби SGML і, що найважливіше, є невеликим і простим у застосуванні.

Усе це схилило нас до застосування принципів TEI для кодування текстових корпусних даних у проекті Українського національного корпусу (УНК). Першим кроком такого застосування стала електронна лексична картотека Інституту української мови НАНУ, про що детальніше йдеться у роботі [3].

У корпусному мовознавстві традиційно текстові корпусні дані розглядають як первинні дані і йдеться про неанотовані довільні тексти природної мови, що мають електронну форму. Кодування первинних даних це по суті описове подання (а) структури електронного тексту: частин, заголовків, абзаців, цитат, речень; (б) незалежно від мови типових текстових елементів, які можуть мати як інтратекстовиий, так і екстратекстовий характер; (в) типових одиниць лінгвального рівня, які виділено у тексті написанням, наприклад: оніми, записані з великої літери, виділені курсивом терміни тощо.

Виходячи із концепції TEI, кодування первинних даних повинно мати структурований характер і передбачатиме кодування [3-5]:

1) глобальної структури первинних даних;
2) типографської розмітки і редакторських правок;
3) одиниць рівня абзацу;
4) одиниць рівня речення.

Коректно закодована глобальна структура первинних даних реалізується через елементи <front>, <group>, <body> і <back>, перший і останній з яких є факультативними, які семантизуються як:

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

Структурно електронний документ корпусу може бути окремим монотекстом, наприклад монографія, роман тощо, або політексом, тобто складатися з окремих творів або їх фрагментів, як наприклад, в антології, і залежно від типу тексту – монотекст / політекст – схема кодування повинна задовольняти наступні моделі:

1) монотекст:

```
<TEI.2>
     <teiHeader> [ інформація електронного заголовка TEI ]  </teiHeader>
     <text>
         <front> [ вступ ...]       </front>
         <body>  [ основна частина ... ]     </body>
         <back>  [ закінчення ...  ] </back>
     </text>
</TEI.2>
```

2) політекст:

```
<TEI.2>
     <teiHeader> [ інформація заголовка об'єднаного тексту ] </teiHeader>
     <text>
         <front> [ втуп 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>
          [ інші тексти або групи текстів ]
        </group>
        <back>  [ закінчення об'єднаного тексту ] </back>
    </text>
</TEI.2>
```

У наведених прикладах чітко простежуємо ієрархічну структуру документа, де текст (<text>) або згруповані кілька текстів в один текстовий документ (<group>) є одиницями найвищого рівня.

Для українського корпусу в межах елементів <text> і <group> пропонуємо виділяти субелементи а) – текстовий розділ в межах текстового документа без підрівнів на взірець <div1>, <div2>, <div3> і т. д., для яких зарезервовано єдиний загальний теґ , б) <p> – абзац і в) <chapter> – частина в межах розділу.

Елемент оперує трьома атрибутами *type*, *n* та *id* з відповідними значеннями:

- *type* – категоризує текст через: BOOK – книга, CHAPTER – глава, POEM – вірш, SONNET – сонет, SPEECH – репліка і SONG – пісня;

Довільне значення атрибута *type*, присвоєне першому з текстових елементів , залишається релевантним для всіх наступних елементів у межах одного елемента <body>.

- *n* – скорочена але зрозуміла назва або номер розділу, які можна використовувати замість ідентифікатора для позначення цього розділу;
- *id* – унікальний ідентифікатор розділу, який можна використовувати для перехресних посилань чи при інших зв'язуваннях елементів; значення кожного атрибута *id* повинно бути унікальним у межах одного документа.

Наприклад:

```
<div id=WN1 n='I' type='book'>
  <div id=WN101 n='I.1' type='chapter'> </div>
  <div id=WN102 n='I.2' type='chapter'> </div>
   ...
</div>
```

Важливим аспектом кодування первинних даних у корпусі є інтерпретація редакторської розмітки, яка подає додаткову інформацію про текст, специфіку лексичних одиниць, можливо, про його автора тощо. Редакторська розмітка як правило полягає у виділенні певних елементів тексту. В українській текстологічній традиції прийнято виділяти головно назви творів, заголовки, підзаголовки, а також цитати і терміни. Можливе також виділення й акцентованих слів та фраз.

Візуально виділення може бути передано різними способами, наприклад, погрубленням і / або курсивом запису малими літерами, погрубленням і / або курсивом запису великими літерами, петітом, розрядкою літер тощо. Застосовуючи принцип TEI, для неінтерпретованого типографського виділення використаємо елемент <hi>, який маркує слово або фразу, що графічно відрізняються 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', 'ЛК', '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. **ЛК** 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=bold>ЛК</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>, де:

- <emph> – виділення емфази;

- <foreign> – виділення запозичення;
- <mentioned> – виділення цитованого або ілюстративного матеріалу;
- <term> – виділення терміна;
- <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, де допустимими є значення: M – для назви монографії; S – назва серії; J – назва журналу; U – для назв неопублікованих матеріалів; A – для назв одиниць, опублікованих як частина більшої; type – класифікує назви відповідно до прийнятої типології через: ABBREVIATED – абревіація, MAIN – основна назва, SUBORDINATE – підзаголовок і назва частин та PARALLEL – альтернативні назви).

Наприклад, виділення терміна-запозичення в уривку:

```
   Документ є міжнародним стандартом на опис розміченого електронного тексту. Точніше,
SGML - це метамова, тобто, засіб формального опису мови, в нашому випадку, мови розмітки
Перш, ніж продовжувати, визначимо ці терміни.
```

```
 <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</foreign>),
    тобто, засіб формального опису мови, в нашому
    випадку, <term type=bold>мови розмітки</term>
    (<foreign lang='en'>markup language</foreign>).</s>
    <s>Перш, ніж продовжувати,визначимо ці терміни.</s>
 </p>
```

Процес кодування первинних даних має багато спільного з процесом редагування у друкарській справі. В обох випадках стоїть завдання зафіксувати стан джерела, всі редакторські виправлення і зміни, запропоновані або внесені до текстового документа у процесі кодування. Збереження первинного вигляду першоджерела і забезпечення зв'язку між джерелом і закодованим текстом є особливо важливим для УНК, оскільки до його складу входитимуть тексти історично різних періодів, з, відповідно, різним правописним узусом, який може і буде відрізнятися від сучасного. А за умови певних редакцій, посилання на першоджерело є передумовою збереження оригінального запису і уникнення фальсифікації фактичного матеріалу. Крім того, явні помилки, які часто зустрічаються в сучасних українськомовних текстах головно офіційно-ділового і розмовного стилів, пропонуємо виправляти і саме виправлення засвідчувати. Загалом, 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, ґ) суржикізми, д) орфографічні помилки, е) семантичні ляпи.

Редагування первинних даних безпосередньо пов'язано з кодуванням редакторських змін, яке також повинно задавати зв'язок поправлених первинних даних з оригіналом і фіксувати відхилення між варіантами тексту. І при кодуванні це доцільно реалізувати за допомогою елементів, детермінованих у TEI: <corr>, <sic>, <orig> і <reg>, де:

- <corr> – засвідчує правильний запис, який в оригіналі наведений з явними помилками.

Цей елемент оперує набором атрибутів, які вказують на (а) вихідну форму з явною помилкою – sic; (б) особу, яка запропонувала і / або внесла виправлення – *recp*; (в) рівень впевненості щодо необхідності конкретного виправлення – *cert*.

- <sic> – містить текст, який доцільно відтворити без змін, незважаючи на його явну некоректність, помилковість чи неточність.

За необхідності в межах цього елемента можна послуговуватися атрибутами елемента <corr>.

- <orig> – запис, зафіксований в оригіналі, можливо навіть помилковий;
- <reg> – виправлений запис з атрибутами *orig* – невиправлений варіант тексту-джерела і resp – особа, відповідальна за виправлення.

Наприклад, уривок з твору Лесі Українки „На полі крові" вимагає наступних коментарів:

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

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 наступне:

```
<text>
  <p>
   <s>Дідок-прочанин іде <reg orig=поз>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, які слід узуповнити, повтори, які є явною помилкою і їх слід зняти, нерозбірливий запис або затерті чи знищені частини тексту тощо. Для цього в TEI передбачено використовувати елементи <add>, <gap>, <del> і <unclear> з відповідною семантикою і атрибутами:

- <add> – текстова одиниця (буква, слово, фраза), узуповнена на місці пропуску.

Елемент <add> оперує атрибутом place – місце вставлених одиниць з одним із можливих значень: INLINE – у рядку, SUPRALINEAR – над рядком, INFRALINEAR – під рядком, LEFT – на лівому полі, RIGHT – на правому полі, TOP – вгорі сторінки, BOTTOM – внизу сторінки.

- <gap> – місце пропуску.

Атрибути цього елемента *desc* – опис опущеного тексту і *resp* – відповідальний за пропуск.

- <del> – містить текст вилученого матеріалу.

Атрибути: *type* – тип вилученого матеріалу відповідно до прийнятої класифікації, *status* – можна використовувати для позначення помилкових видалень і *hand* – вказує на особу, яка вилучила матеріал.

- <unclear> – текстова одиниця, яка не піддається ідентифікації через технічні ушкодження тексту.

Атрибути: *reason* – подає причину складності ідентифікації запису і *resp* – вказує особу, яка розшифрувала запис.

Наприклад, якщо оригінал електронного тексту:

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

Концепція схеми кодування первинних даних TEI є абзацоцентричною. І наступну групу елементів розглядаємо як елементи рівня абзацу. Загалом на абзацному рівні прийнято розрізняти (а) елементи власне абзацу (б) так звані міжрівневі елементи, тобто структурні одиниці, які можуть з'являтися як у межах абзацу, так і за його межами або самі становити абзац і (в) фразові елементи. А структурними елементами первинних даних у межах текстового абзацу можуть бути 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. В Українському національному корпусі на рівні абзацу в первинних даних пропонуємо кодувати (1) цитати, (2) віршові рядки, (3) списки, (4) таблиці, (5) адреси і (6) речення.

Так, цілісний текстовий абзац кодується елементом <p>, в межах якого для кодування детермінованих вище одиниць призначені теги <q>, <l>, <lg>, <sp>, <speaker>, <stage>, <list>, <table>, <addres> і <s>.

Схема TEI розрізняє п'ять типів цитованого матеріалу, який може з'являтися в межах абзацу: 1) інтратекстова цитата, 2) інтертекстова цитата, 3) цитата з посиланням, 4) ілюстративний матеріал, поданий як цитата і 5) текстовий уривок, уведений в контекст словами „*так би мовити*", „*як стверджує X*" etc. Для кожного з цих типів цитованого матеріалу передбачені відповідні елементи кодування:

- <q> (інтратекстова цитата) ідентифікує мову автора в межах власного твору, подану як цитату, і оперує атрибутами type – уточнює тип цитованого матеріалу (репліка, роздуми, експлікандум тощо) через: SPOKEN – пряма мова і THOUGHT – роздуми, внутрішні діалоги; who – ідентифікує мовця; direct – вказівка непряму мову;
- <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 конструктів на кшталт „так би мовити", „за словами Пана Х", „як сказав У".

Наприклад, у первинних даних УНК не передбачено ідентифікації усіх п'яти типів цитат. Усі ці типи пропонуємо кодувати елементом інтертекстової цитати <quote>:

На дві 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>
```

Організація віршового і драматичного текстів формально відрізняється від прозового тексту, що необхідно подати у процесі кодування первинних даних. У TEI передбачено ідентифікувати віршовий і драматичний текст як у межах прозового тексту, так і індивідуально, та кодувати їх за схемою TEI такими тегами <l>, <lg>, <sp>, <speaker>, <stage>, де:

- <l> – подає рядок віршового тексту, а метричну завершеність / незавершеність конкретного рядка специфікує атрибут part;
- <lg> – кодує групу віршових рядків, які формально становлять цілісну одиницю;
- <sp> – ідентифікує індивідуальне мовлення, організоване як вірш, в межах прозового тексту, з вказівкою на мовця через атрибут *who*;
- <speaker> – елемент, призначений для забезпечення інформації про власну назву мовця / мовців у драматичному тексті;
- <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 в Українському національному корпусі. Тому схема кодування первинних даних повинна обов'язково забезпечувати ресурси для машинного подання такого тексту. TEI трактує список як впорядковану / невпорядковану послідовність текстових одиниць або глосарій. Перед кожною з одиниць списку може стояти певна мітка – цифра, літера – а у глосарії такою міткою є термін, що визначається. Схема кодування списку оперує набором елементів <list>, <item>, <label>, <head>, <headLabel>, <headItem>, які оперують або глобальними атрибутами: *id* – унікальний ідентифікатор елемента, який має значення ідентифікатора, *n* – номер або інша позначка, не обов'язково унікальна у межах корпусу, конкретного заголовкового елемента, *lang* – ідентифікує мову теґового запису через дволітерний код з ISO 639 чи трилітерний код з ISO 639-2, чи один з розширень коду країни з ISO 3166; або атрибутами *type*, *ordered*, *bulleted*, *gloss*.

Семантика тегів кодування таблиці наступна:

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

Елемент <list> уточнює власну семантику за допомогою атрибутів *type*, який задає формальні параметри списку і його коректними значеннями можуть бути *ordered* та *bulleted* – відповідно для списків з цифровою, буквеною чи символьною нумерацією; *gloss* – для списків, що складаються з набору технічних термінів, кожний з яких відзначено елементом <label> і супроводжено дефініцією, маркованою елементом <item>, і *simple* – для списків, пункти яких не пронумеровані чи відзначені в інший спосіб.

- <item> – структурний елемент списку;
- <label> – мітка, зв'язана зі структурним елементом списку; в глосаріях маркує тлумачений термін;
- <head> – довільний заголовок або підпис списку чи його частини;
- <headLabel> – підпис мітки або терміна-мітки в глосарії чи структурному елементі списку;
- <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, а кожну окрему одиницю списку слід кодувати окремим елементом <item>. Перед першим з елементів <item> можна але не обов'язково розмістити елемент <head> із заголовком списку. Залежно від вимог оброблення списку, нумерація в ньому може бути а) пропущена, б) визначена за допомогою атрибута *n*, або в) розмічена теґом <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>
```

Зауважимо, що схема кодування списку TEI не дозволяє одному і тому самому спискові використовувати різні стилі кодування одночасно. Але довільний список може мати як завгодно глибоко вкладену структуру.

Для списку літератури, тобто бібліографії, TEI резервує елемент <listBibl> і розмічає її у системі бібліографічного посилання. А якщо список має складну внутрішню структуру, його доцільно інтерпретувати як таблицю.

Таблиця є двомірним способом формалізованого подання інформації. Інтерпретація таблиць є складним завданням для будь-якої системи оброблення текстів, але оскільки такий тип формалізації знань поширений, то TEI забезпечує методи кодування таблиці, визначаючи елементи кодування та їх атрибути:

- <table> – текст, візуалізований як таблиця, тобто складається з рядків і стовпців:

аналізований елемент оперує двома атрибутами 1) *rows* – вказівка на кількість рядків у таблиці, 2) *cols* – вказівка на кількість стовпців у кожному рядку таблиці.

- <row> – рядок таблиці з атрибутами: *role* – вказівка на тип інформації, збереженої у гнізді рядка через *label* – для міток чи описової інформації і *data* – для реальних значень даних;
- <cell> – гніздо[1] таблиці:

атрибутами цього елемента є *ole* – вказівка на тип інформації, збереженої у гнізді через детерміновані вище: *label* і data, *cols* – вказівка на кількість стовпців, що займає конкретне гніздо та rows – вказівка на кількість рядків, що займає конкретне гніздо.

Не дивлячись на те, що схема TEI відносить адресу до розряду міжрівневих елементів, в українськомовній аплікації принципів TEI, адресу інтерпретуємо як елемент абзацного рівня і кодування адрес різного типу – класична, електронна тощо – здійснюємо за допомогою теґів <address> – поштова або інша адреса і <addrLine> – окремий рядок поштової або іншої адреси. Наприклад:

```
<address>
<addrLine>Інститут української мови</addrLine>
<addrLine>Грушевського, 4</addrLine>
<addrLine>Київ, 01001</addrLine>
<addrLine>Україна</addrLine>
</address>
```

Кодування адреси можна також деталізувати, увівши елемент <name> – власна назва, наприклад:

```
<address>
<addrLine>Інститут української мови</addrLine>
<addrLine>Грушевського, 4</addrLine>
<addrLine><name type=city>Київ</name>, 01001</addrLine>
<addrLine><name type>=country>Україна</name></addrLine>
</address>
```

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 на обов'язкове виділення цього елемента. Маркування орфографічного речення передбачено реалізувати шляхом застосування теґа <s> – текстові уривки, які відповідають вимогам орфографічного 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 всю 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>
```

Кодуючи первинні дані в Українському національному корпусі, ми зіткнемося з проблемою подання а) слів, які записують згідно з принципами українського правопису з великої ↔ малої літери, б) абревіатур, в) цифр, оформлених бквенно і літерно, і в) пунктуаційних символів.

Вживання великої і малої літери згідно з принципами українського правопису торкається опозиції „онім – апелятив". Диференційною ознакою подання оніма в українськомовних текстових даних є його запис через початкову велику літеру. У первинних даних в межах групи слів-онімів інтерпретують класичні оніми, детерміновані системою мови і маркують їх теґом <name> –

[1] У науковій літературі можна зустріти синонімічний термін чарунка.

довільна власна назва, який оперує атрибутом *type*, призначення якого типологізувати кодований онім.

Проблема стандартного подання онімів в УНК головно виникатиме, коли йтиметься про антропоніми, які в українській традиції можуть бути двокомпонентними або трикомпонентними: 1) ім'я + прізвище і 2) прізвище + ім'я + по-батькові. З метою досягнення антропонімної уніфікованості TEI в межах елемента <name> забезпечує два додаткові атрибути 1) *key* – альтернативний ідентифікатор для об'єкта з певним онімом, подібний до ключа запису в базі даних, і 2) *reg* – уніфікована або виправлена онімна форма.

Валерій Шевчук розпочинає свій твір „Три листки за вікном" цитатою з **Г.Сковороди** „Світ неситий, коли не задовольняє. Вічність несита, 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>Г.Сковороди</name> <quote>Світ неситий, коли не задовольняє. Вічність несита,
коли не завдає жалю... 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 абревіатури 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.

У рукописах, де як правило багато різних скорочень, атрибут *type* можна використовувати для виділення типів цих скорочень за їхньою функцією, а атрибут *expan* – для визначення повної форми скорочень.

Наприклад, фрагмент тексту '*США – це супердержава сучасного світу*' можна закодувати наступним чином:

```
<p>
 <s><abbr type=geogr>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 і запис відповідати національній традиції. Не залежно від принципу фіксації числової інформації, диференційна ознака якої детермінована десигнатом, її кодування можна здійснювати за допомогою елементів <num>, <date> і <time>, де:

- <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.

- <date> – дата у довільному форматі запису.

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

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

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

Крім того, неповні дати або час можна подати з пропусками або через діапазон дат або інтервал часу, а якщо одна з меж відома, доцільно використати атрибут *exact*.

Наприклад:

```
<num value='33'>xxxiii</num>
<num type=cardinal value='21'>двадцять один</num>
<num type=percentage value='10'>десять процентів</num>
<num type=percentage value='10'>10%</num>
<num type=ordinal value='5'>5ий</num>
<date value='1980-02-21'>21 люте 1980</date>
<date value='1990'>1990</date>
<date value='1990-09'>Вересень 1990</date>
Встановлений <date value='1977-06-12'>дванадцятого дня
червня.</date>
<l>спеціально, коли було дев'ять нижче нуля
<l>і <time value='15:00'> третя година після обіду</time>
```

Кодування пунктуаційних символів у корпусі може мати факультативний характер. Винятком повинні бути дані для синтаксичних досліджень, у яких слід розмітити крапки, знаки питання, знаки оклику, дефіси, тире, лапки й апострофи. Зауважимо, що семантика розділових

знаків залежно від позиції і відступів перед і після символу є різною. Так, крапка може експлікувати кінець речення, абревіатуру, еліпсу (в три крапці) або візуалізувати місце для дати чи підпису. Багатозначність крапки прийнято усувати шляхом інтерпретації пре- і постпозиційного оточення цього символа, йдеться про наявність / відсутність відступів і / або інших символів.

Знак питання і знак оклику традиційно позначають кінець речення. Але можливі випадки використання цих знаків у середині речення з метою акцентування уваги на певній одиниці речення. Неоднозначність використання цих знаків знімається шляхом інтерпретації маркування. Знак оклику або знак питання в межах s-одиниці може мати суміжні знаки – дужки, лапки тощо, або відділятися пробілами від інших знаків, а в кінці речення як правило таких суміжних символів немає і пробіл можливий лише в постпозиції щодо знаку питання чи оклику.

Дефіс може позначати постійний знак переносу в слові. Якщо синтагматичний порядок усіх елементів аналізованого речення в машиночитаному тексті відрізняється від тотожного в оригіналі, редактор повинен або усунути несуттєвий знак переносу, або замінити його на відповідний символ. Але не залежно від оброблення дефісу в конкретному кодуванні, стратегія оброблення декларується в електронному заголовку TEI через елемент <hyphenation>.

Тире дещо детальніше інтерпретовано у схемі кодування TEI. Цей пунктуаційний знак забезпечено стандартною назвою у видавничому наборі символів ISOpub і детерміновано в ISO 8879 як: **mDash**, **nDash** і **dash**, залежно від конкретної семантики тире.

Лапки теґуються двояко: як <q>, або як <quote>, де перший із елементів маркує власне символ лапок, а другий – саму цитату.

Апостроф згідно з вимогами TEI повинен відрізнятися від лапок. Найоптимальніше реалізовувати розрізнення цих двох надрядкових символьних позначень шляхом маркування лапок або їх використання теґами <q> або <quote>, а апостроф не маркувати взагалі.

Висновки. Навіть з короткого огляду принципів кодування первинних даних і маркаційних ресурсів TEI, з'ясовується, що довільний текст природної мови, розмічений теґами, детермінованими TEI, перетворюється на стандартний об'єкт IT-середовища з можливістю багатократного та різноаспектного застосування. Крім того, підходи TEI до подання текстового ресурсу забезпечують (а) уніфіковане і стандартне оброблення шляхом застосування описової розмітки лінгвальних елементів різного рівня, зокрема тіла тексту, заголовка, абзацу, таблиці, списку, слова тощо, (б) збереження зв'язку між оригінальними і закодованими даними, (в) нескладну реалізацію кодування, причому без спеціалізованого програмного забезпечення, і (г) можливість роботи з будь-яким SGML-сумісним ресурсом.

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.
3. *Демська-Кульчицкая О.М., Перевозчикова О.Л., Сичкаренко В.А.* Организация и ведение лексической картотеки украинского языка // Проблемы программирования. – 2002. - №1-2. – С.512-516.
4. *Перевозчикова О.Л.* Сучасні інформаційні технології. – К.: Інститут економіки і права "КРОК", 2000. – 124 с.
5. *Перевозчикова О.Л.* Стандартизація і сертифікація інформаційних технологій. – К.: Університет економіки і права „КРОК". – 2003. – 215 с.