УДК **811.161.2-116**

Демська-Кульчицька О. М.

ЗАГОЛОВОК ЯК ОБОВ'ЯЗКОВИЙ СТРУКТУРНИЙ ЕЛЕМЕНТ ЛІНГВІСТИЧНОГО КОРПУСУ

У статті йдеться про необхідність розвитку нового лінгвістичного напряму в лінгвоукраїністиці - корпусної лінгвістики, який мав би починатися з побудови Національного корпусу української мови. Проаналізовано можливості використання Принципів TEI до створення електронного корпусного і текстового заголовка й показано прийом аплікації моделі заголовка TEI у створенні електронного заголовка, який є обов'язковим структурним елементом корпусу текстів.

Сучасна лінгвістика уже не ставить під сумнів існування в її парадигмі корпусної лінгвістики - напряму, завданням якого є розроблення теоретичних засад і практичних прийомів побудови, машинного опрацювання та експлуатації емпіричних лінгвальних даних, оформлених як корпус текстів. Впровадження цього нового напряму в лінгвоукраїністику є важливим завданням і передусім передбачає теоретичне обгрунтування та практичне створення *Національного корпусу української мови (НКУМ)* - перетвореної на електронну форму, системно організованої та програмно обробленої вибірки текстів української мови, репрезентативних для всіх як історичних, так і географічних варіантів та форм її існування.

Детермінативними ознаками корпусної моделі організації емпіричного матеріалу є можливість його багатократного та різноаспектного використання, починаючи від лінгвістичного дослідження і закінчуючи технологічними застосуваннями. Щодо лінгвістичного аналізу, то він передбачає академічні лінгвістичні дослідження різних рівнів мовної системи: лексики, морфології, синтаксису тощо, а також методику викладання / навчання / вивчення мови як рідної та як іноземної. Натомість у технологічному застосуванні йдеться про використання корпусу з метою побудови машинної мовної моделі як основи для розробок у галузі інформаційних технологій, створення програм автоматичного розпізнавання і синтезу мовлення, забезпечення автоматичних методів перетворення текстової інформації, лінгвістичної підтримки автоматичних систем управління.

Багатократність використання корпусного ресурсу та різноаспектність запитів до нього зумовлена електронною формою існування корпусу і особливо стандартністю його подання. Важливо, що умова електронної форми існування недостатня для визнання збірки електронних текстів корпусом. Необхідною передумовою корпусної кваліфікації електронного об'єкта є дотримання у процесі його створення Стандартів кодування корпусу, в основі яких лежить принцип мовної незалежності.

Стандартність подання корпусу передбачає обов'язкову наявність у його структурі **електронного заголовка** і тіла документа чи власне закодованого тексту.

Традиційно під **електронним заголовком** розуміють початкову частину електронного документа, в якій через систему спеціальних елементів подано корпусну або текстову метаінформацію. Як правило, метаінформація охоплює бібліографічні, екстра- та інтралінгвістичні текстові дані, опис методик кодування, теговий набір, перегляд і верифікацію процесу кодування тощо. По суті, електронний заголовок є аналогом титульної сторінки класичного друкованого видання.

У корпусі функціонують електронні заголовки двох типів: електронний заголовок до усього корпусу, чи **корпусний заголовок,** і електронний заголовок до конкретного текстового файла, який входить до складу корпусу, чи текстовий **заголовок.** Відмінність між цими двома типами електронних заголовків полягає у тому, що корпусний заголовок забезпечує метаінформацію про корпусний проект загалом, а текстовий - про конкретний текст чи фрагмент тексту, збережений у файлі.

Уперше концепція електронного заголовка і засоби кодування інформації у ньому були розроблені в межах проекту *Ініціативи кодування тексту* (TEI [1]). Сьогодні у корпусній лінгвістиці

[1] TEI (Text Encoding Initiative) - ініціатива кодування тексту є міжнародним і міждисциплінарним стандартом подання усіх типів текстів, функціональних у бібліотечній, музейній, видавничій справах та мовознавстві, шляхом використання максимально виразної і мінімально застарілої схеми кодування. Призначення TEI - забезпечити коректний обмін текстової інформації, яка має електронний вигляд та засоби експлікації архітектоніки тексту, з метою спрощення його оброблення програмними засобами. Загалом на сьогодні 88 проектів реалізовано або реалізовують, використовуючи Принципи TEL

до написання електронного заголовка переважно застосовують SGA/L-Версіїо *Стандарту кодування корпусу (Corpus Encoding Standard)* або наступну XML-версію цього ж стандарту *(XCES),* базовану на TEI-схемі. Суттєвої відмінності між цими стандартами немає, і такий підхід реалізований у електронних заголовках *Британського національного корпусу, Польського корпусу Інституту основ програмування Національної академії наук, Великого корпусу російської мови* (БОКР) та ряді інших.[2] TEI-схему пропонуємо реалізувати і в електронному заголовкові *Національного корпусу української мови.*

Електронний заголовок за TEI-схемою задають елементом <teiHeader>. Цей елемент складається з чотирьох частин, кожна з яких об'єднує ряд піделементів і подає інформацію про:

а) <fileDesc> - бібліографію електронних корпусних текстів;

б) <profileDesc> - характерні екстра- та інтралінгвістичні параметри тексту, наприклад, інформацію про стилістично-жанрову специфіку тексту, мову/субмову тощо;

в) <encodingDesc> - принципи, структуру та методи кодування даних;

г) <revisionDesc> - історію модифікації оброблення даних.

Кожен елемент найвищого рівня оперує рядом субелементів, призначених для деталізації інформації. Технічно електронний заголовок записують так:

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title> .....</title>
      <autor> ... </autor>
    </titleStmt>
    <extent>
      <count unites=word> ...</wordcount>
      <bytecount units="bytes"> ... </bytecount>
    </extent>
    <respStmt>
      <resp> ... </resp>
      <name> ... </name>
    </respStmt>
    <publicationStmt>
      <publisher> ... </publisher>
      <pubPlace> ... </pubPlace>
      <address> ... </address>
      <date> ... </date>
      <idno type=ISBN> ... </idno>
    </publicationStmt>
    <sourceDesc>
      <monogr>
        <title> ... </title>
        <autor> ... </autor>
        <data> ... </data>
          <address type=web page> ... </address>
      </monogr>
    </sourceDesc>
  </fileDesc>
  <profileDesc>
    <creation>
      <date> ... </date>
    </creation>
    <language id=ua> ... </language>
    <textClass>
      <keywords>
        <term> ... </term>
      </keywords>
    </textClass>
    </profileDesc>
  </profileDesc>
  <encodingDesc>
    <projDesc> ... </projdesc>
    <sampDesc> ... </sampDesc>
      <editlDecl>
        <conformance level= > ... </conformance>
    </editDecl>
    <tagsDecl>
        <tagUsage gi=text occurs=1>
        <tagUsage gi=body occurs=1>
        <tagUsage gi=p occurs=12>
        <tagUsage gi=hi occurs=6>
      </tagsDecl>
    </encodingDesc>
    <revisionDesc>
      <change> ... </change>
        <date> ... </date>
        <respStmt>
          <name> ... </name>
        </respStmt>
    </revisionDesc>
  </teiHeader>
```

У наведеній мінімалізованій моделі електронного заголовка TEI першим структурним елементом є <füeDesc> (буквально 'опис файла'). Це обов'язковий елемент електронного заголовка і його формують під елементи:

- <titleStmt> - бібліографічна інформація про текст задана через <title> - заголовок або назва твору/праці, <autor> - автор;

-<extend> - обсяг кодованого тексту [3];

-<respStmt> - вказівка на відповідальність за створення електронних даних: тип відповідальності - <resp> і власна назва відповідального - <name>;

- <publicationStmt> - бібліографічна інформація про видання, яку задають елементом <рш> lisher> - назва організації, відповідальної за публікацію чи дистрибуцію бібліографічної одиниці, тобто видавництво або видавець, і деталізовано <pubPlace> - місце функціонування видавництва/видавця, <address> - поштова чи будь-яка інша

[2] Див.: http//www.cs.vassar.cdu

[3] У елементі <cxtcnd> інформацію можна подати через «w» -

адреса видавництва/видавця, <date> - дата публікації описуваної одиниці, <idno> — довільний стандартний або нестандартний номер, використовуваний для ідентифікації бібліографічної одиниці з атрибутом type, який вказує на тип бібліографічної класифікації, наприклад, індекс ISBN;

- <sourceDesc> - опис першоджерела, оформлений за допомогою субелементів відповідно: <monogr> - монографія, <analytic> - збірник, а також <title>, <autor>[4], <date>, <publisher>, <address>.

Наступним структурним елементом електронного заголовка є <profileDesc> ('опис профілю тексту') - структурований опис лінгвістичних та екстралінгвістичних параметрів кодованого тексту. У межах цього елемента згідно з TEI-схемою можливі субелементи <creation> - інформація про місце і час створення тексту[5], <langUsage> - вказує на мову, субмову, діалект etc. тексту, <textClass> - групує інформацію про типологію тексту[6] та ін.

У межах опису профілю можна також подавати інформацію про типологію текстів, для чого передбачено елемент <textDesc>. Інформацію у цей елемент уводять через субелементи: <chanal> - вказівка на тип комунікаційного каналу (усне живе мовлення, усне телефонне мовлення, електронний лист, факс тощо); <domain> - сфера застосування (освіта, наука, інтерв'ю тощо). Набір елементів у <textDesc> найчастіше є індивідуальним у кожному конкретному корпусі чи корпусному проекті. Так, у *Національному корпусі української мови* в межах аналізованого елемента, крім вказівки на тип комунікаційного каналу, передбачено інформацію про стиль і жанр текстів.

Проблема стилістично-жанрової класифікації у корпусі національного типу є однією з неоднозначних, для якої на сьогодні корпусна лінгвістика не пропонує цілковитого вирішення. Лише як рекомендовані пропоновано використовувати елементи: <usg type=«style»>, <prepared nes> та <interaction>.

У проекті *НКУМ* пропонуємо замінити ці елементи і замість них використати такі: <style> - стильова характеристика тексту, деталізуючи її в атрибутах pen - художній стиль, science - науковий, busin - офіційно-діловий, геї - конфесійний, epist - епістолярний, speak - розмовний; <genre> - жанрова специфіка тексту з prose - проза, poetry - поезія і drama - драма.

У наступному елементі <encodingDesc> ('опис кодування') передбачено подавати інформацію про принципи, структуру і методи кодування корпусного тексту(ів). Цю інформацію вводять за допомогою елементів: <projDesc> - довідка про проект, в межах якого реалізовано побудову конкретного корпусу, та процедури формування даних, <sampDecl> - описує принципи відбору фактичного матеріалу до корпусу, <editDecl> - визначає засади редагування, застосовані у процесі кодування первинних даних, <tagsDecl> - уводить набір тегів кодування. Зауважимо, що інформацію в елементах <projDesc>, <sampDecl> і <editDecl> доцільно оформлювати як традиційний текст.

Останнім структурним елементом електронного заголовка в моделі електронного заголовка TEI є елемент <revisionDesc>, у якому прийнято подавати інформацію про історію модифікацій, тобто ця частина по суті є протоколом змін, внесених у процесі кодування корпусу. Дані у цьому елементі вводять через субелементи <change> - вказівка на зміст змін(и), <date> - дата внесення змін(и), <respStmt> - відповідальність за зміни(у) і відповідно <resp> - тип відповідальності та <name> - ім'я/назва відповідального.

Отже, схематично TEI заголовок із введеною інформацією має відповідати поданню, наприклад:

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title> Субстантиви в українській мові </title>
      <autor> Григоренко Оксана Степанівна </autor>
    </titleStmt>
    <extent>
      <count unites=word> 3 543 </wordcount>
      <bytecount units="Kbytes"> 134 </bytecount>
    </extent>
    <respStmt>
      <resp>
        редакторські правки пунктуаційних знаків у 5 і 46 реченнях згідно з чинним Правописом української мови
      </resp>
      <name> Лебеденчук Катерина Василівна </name>
    </respStmt>
    <publicationStmt>
```

[4] У елементі <autor> специфіку авторства задають через: «mono» - один автор (дані «на промовчання»), «poli» - група авторів і «corporate» - авторство тексту належить кільком організаціям, наприклад, як *Декларація прав людини*.

[5] У <creation> як правило використовують субелементи; <date> - дата, <rs> - рядок відсилання для вказівки на місце створення тексту, <person> - характеристика автора тексту.

[6] Фактично цей елемент задає інформацію про таксономію тексту, реалізуючи це або через елемент <keyword> - ключові слова, які по суті співвідносять конкретний текст з певною предметною галуззю, або через елемент <classCode>, у якому задають стандартну класифікацію на взірець УДК, або ж через елемент <catRef> - посилання на спеціальну, розроблену для конкретного корпусу, класифікацію текстів.

```
        <publisher> Відродження </publisher>
        <pubPlace> Київ </pubPlace>
        <address> Васильківська, 27 </address>
        <date> 2003 </date>
        <idno type=ISBN> 23-2341-234-0 </idno>
    </publicationStmt>
    <sourceDesc>
            <analytic>
                <title> Граматичні студії </title>
                <edit> Варнак Петро Іванович
                                        </edit>
            </analytic>
    </sourceDesc>
</fileDesc>
<profileDesc>
    <creation>
        <date> 24.05.04 </date>
    </creation>
    <language id=ua> українська </language>
    <textClass>
        <keywords>
            <term id=1> граматика </term>
            <term id=2> субстантив </term>
        </keywords>
    </textClass>
    </profileDesc>
</profileDesc>
<encodingDesc>
    <projDesc>
        Спільний проект Інституту української мови
        та Інституту кібернетики НАН України, роз-
        початий у 2002 році
    </projdesc>
    <sampDesc> Випадкова вибірка </sampDesc>
        <editlDecl>
            <conformance level= > 3 </conformance>
        </editDecl>
```

```
        <tagsDecl>
            <tagUsage gi=text occurs=1>
            <tagUsage gi=body occurs=1>
            <tagUsage gi=p occurs=12>
            <tagUsage gi=hi occurs=6>
        </tagsDecl>
    </encodingDesc>
    <revisionDesc>
        <change> немає </change>
            <date> 02.08.04 </date>
        <respStmt>
            <name> Процків Сидір Миколайо-
                                    вич </name>
        </respStmt>
    </revisionDesc>
</teiHeader>
```

Такий формат забезпечує інформацію, достатню для коректних наукових застосувань та висновків, що важливо як з огляду лінгвістичних результатів, так і програмних застосувань. Наприклад, аналіз відхилень у парадигмі відмінювання можна кваліфікувати як помилку за умови відсутності інформації про дату видання та правописні норми, використані у конкретних корпусних даних. Крім того, TEI-схема подання електронного заголовка забезпечує коректне співвідношення між набором засобів кодування первинних даних і метаінформацією, а збудований таким чином електронний заголовок у корпусі довільної природної мови перетворює останній на стандартний корпусний об'єкт, що дозволяє використовувати для роботи з таким ресурсом наявні корпусні пошукові засоби.

1. *IdeN.* Corpus Encoding Standard.- http://lpl.univ.aix.fr/projccts/multext/CES, 2000.

2. *Sperberg-McQueen C. M., Bwnard L.* Guidelines for Electronic Text Encoding and Interchange- http://www.hcu.ox.ac.uk/TEI/P4X/indcx.html, 2001.

O. Demska-Kulchytska

HEADER AS OBLIGATORY STRUCTURAL ELEMENT OF LINGUISTIC CORPUS

The article deals with the necessity of development the direction of Corpus Linguistics in Ukrainian linguistic science and creation the Ukrainian National Corpus. It is analyzed the possibility of application the TEI Guidelines for Electronic Text Encoding and Interchange to corpora creation, and also proposed the example of TEI-header as obligatory element of the texts corpus.