ТЕГУВАННЯ ДАТ В НАЦІОНАЛЬНОМУ КОРПУСІ УКРАЇНСЬКОЇ МОВИ

© Оксана Тищенко, 2007
к. філол. н., Інститут української мови НАНУ (Київ)
УДК 811.161.2'374.72'22

У статті йдеться про кодування дат у текстах Національного корпусу української мови. Розглянуто власне поняття дати, способи її репрезентації в текстах різних стилів та можливі варіанти тегування.

Створення корпусів мови як специфічно організованих текстових множин зумовлено тим, що певні дослідження в сучасних умовах інформаційного суспільства можна здійснювати лише на основі значного за обсягом мовного матеріалу, послуговуючись найновішими методами й технологіями в дослідженнях природних мов. Світовий досвід у цій галузі представлений численною кількістю репрезентативних корпусів як національних мов, так і мов окремих авторів чи функціональних стилів, діалектів та ін. [2].

Одним з наріжних питань у корпусній лінгвістиці є проблема кодування первинних даних та лінгвістичне анотування, тобто введення формалізованої лінгвістичної інформації в електронний текст (тегування). У Національному корпусі української мови (НКУМ) прийнято формат подання, що відповідає принципам TEI [1].

Існує кілька типів анотації – довільної лінгвістичної інформації про лінгвально релевантні одиниці текстових даних і наявність такої інформації в тексті – семантична, анафорична, дискурсна, прагматична, морфологічна тощо [1]. Розглянемо проблему семантичного анотування, зокрема представлення в корпусі дат і часу.

Дата – це календарний час якої-небудь події; позначка, що вказує час (рік, місяць, число) написання чого-небудь (листа, статті тощо) [4].

Отже, анотуванню підлягає позначка в тексті, що вказує на календарний час будь-якої події. Для визначення змісту та меж позначки, що підлягає анотуванню, а також змісту самої анотації дат в НКУМ використовуємо дані теоретичних джерел, передусім [1: 164–165; 3]. О. М. Демська-Кульчицька, закладаючи основи Національного корпусу української мови, розглядає кодування дат у межах кодування первинних даних – як число, тобто пріоритет надано не даті, а числовому позначенню величин, зокрема дат. Так, автор пропонує використовувати теги:

<num> – довільно записане число з атрибутами *type*, який експлікує тип числового значення (дріб, порядковий числівник тощо), і *value* – стандартне подання значення числа;

<date> – дата в довільному форматі запису за атрибутами *calendar,* що визначає систему числення чи календар, за яким подано дату, *value* – стандартне подання дати у форматі «рік–місяць–день»; цей атрибут визначає стандартний запис через Міжнародну Організацію зі Стандартизації (МОС) 8601 (ISO 8601). МОС 8601 описує значну кількість форматів дати й часу, наприклад, основний формат (без пунктуації) та розгорнутий (з використанням пунктуації), що дає можливість опускати певні елементи, як-от століття може бути представлене лише двома цифрами – 01.12.98, 13.01.07;

<time> – часова інформація в межах доби в довільному форматі запису:

<date value='1980-02-21'>21 лютого 1980</date>
<date value='1980'>1980</date>
<date value='12.06'>дванадцятого червня</date>
<time value='15:00'>о третій дня</time >
<s>Треба зауважити, що темпи зростання економіки США <date value='2003'>в 2003 році</date> склали 8,3%.</>

Такі анотації цілком задовільні, коли дати й час в анотованому тексті подані чітко й точно. Їх можна розширити, звернувшись до подання часу за МОС 8601, що окрім року (YY або YYYY), місяця (MM), дня (DD) пропонує подання:

– секунд та десяткових часток секунд: ss.s;

– годин і хвилин (hh:mm), виражених як у скоординованому середньому гринвіцькому часі (в UTS) зі спеціальним вказівником UTS («Z»), так і виражених у форматі місцевого часу, з поданням часового поясу в годинах та хвилинах: YYYY-MM-DDThh:mm:ss.sTZD, де TZD вказує на часових пояс (Z ± hh:mm:ss.s), напр.: 2007-09-16T19:28:30.4+02:00.

Однак крім точно вказаного моменту подій (рік, місяць, день, години, хвилини) у текстах трапляються приблизні вказівки, так звані неточні дати, напр., *у другій половині дня, на початку вересня, в кінці року, середина XX ст., з вересня до грудня, за кілька років до проголошення незалежності* тощо. У цьому разі для анотування пропонуємо розрізнювати **точні** й **неточні,** зокрема **оказіональні** (події, виражені в тексті як, наприклад, назва свята, якого-небудь відомого випадку, історичного періоду тощо), **абсолютні** та **відносні** дати й час, а також використовувати атрибут *certainty*, який вказує на ступінь точності, з якою подано дату.

Абсолютна часова анотація містить такі елементи або їхню послідовність: **<day>, <week>, <month>, <year>, <second>, <minute>, <hore>, <occasion>** (останній елемент вживають для анотування оказіональних дат).

Додатково можливе детальніше анотування дат, для чого пропонується тег <**dateStruct**> – містить внутрішнє структуроване представлення, напр., для абсолютної дати, <**timeStruct**> – містить внутрішнє структуроване представлення часу (за джерелом [3]):

Мінімальний обсяг анотації дати й часу в тексті, який вказує тільки на їхню наявність:

<s>Як відомо, <date>у четвер</date> його як свідка було викликано повісткою, підписаною слідчим у особливо важливих справах, керівником слідчої групи у справі Гонгадзе Грищенком.</>

<s>Ті американці, які побували у нас в Полтаві <date>в жовтні 2003 року</date>, більше всього ділились враженнями саме про наш чорнозем. </>

<s><date>11 листопада</date> до редакції зателефонували з виробничого відділу видавництва «Донеччина». </>

<s>Тоді наступного дня, <date>12 листопада</date>, <time>о 9 год. 40 хв.</time> до приміщення, де ще кілька днів тому перебувала редакція «Острова», прибули молоді люди бритоголової зовнішності та вражаючих габаритів.</>

<s>Рішенням господарського суду «Т» області <date>від 25 липня 2003 року</date>, що прийняте суддею, позовні вимоги задоволені у повному обсязі. </>

<s>Незважаючи на драконівські закони царату <date>від 1863</date> та <date>1876 р.</date> щодо української мови, розвиток української культури піднявся на новий рівень.</>

Представлення значення дати:

<s>Ті американці, які побували у нас в Полтаві <date value='2003-10'>в жовтні 2003 року </date>, більше всього ділились враженнями саме про наш чорнозем. </>

Представлення типу вираження дати:

<s>Як відомо, у <date type='name'>четвер</date> його як свідка було викликано повісткою, підписаною слідчим у особливо важливих справах, керівником слідчої групи у справі Гонгадзе Грищенком.</>

Структуроване представлення дати:

<s> Ті американці, які побували у нас в Полтаві
<dateStruct value='2003-10'>
<month type='name' value='–-10'> в жовтні</month>
<year type='num' value='2003–-'>2003 року</year>
</dateStruct >,
більше всього ділились враженнями саме про наш чорнозем.</>

<**dateRange**> – містить дві абсолютні дати або інше визначення, що має значення певного періоду часу (діапазону), містить атрибути *from* – вказує на

початкову точку дати, *to* – вказує на кінцеву точку дати, *exact* вказує на точність приписаних значень:

<s>Населення Лівобережжя підтримало визвольний бунт польських землевласників

<dateRange from='1830' to='1831'>1830–1831 років </dateRange>.</>

<s>Українському національному культурному відродженню сприяло .. заснування популярної газети «Український вісник» <dateRange from='1816' to='1863'>1816–1863</dateRange>.</>

<s>Якщо темпи росту ринку лізингу залишаться такими як

<dateRange from='2005' to='2006'> у 2005 – 2006 роках </dateRange>,

то вартість укладених лізингових угод на кінець року збільшиться з 423 до 660 млн. доларів США. </>

<**timeRange**> – так само містить дві вказівки на час, представлений у стандартній формі, або інше визначення, що вказує на певний проміжок часу (діапазон).

Відносна часова анотація описує дату або час відносно іншого (абсолютного) часового моменту, містить такі елементи:

<**distanse**> – вказує на часовий відтинок, що відмежовує анотовану подію від певної дати, відносно якої згадується;

exact – вказує на ступінь точності, з якою подано часову відстань;

<**offset**> – частина часової анотації, що означає напрям зміщення анотованої події відносно певної дати.

<s><dateStruct value='11-11'>

<day type='num'>11</day>

<month type='name'> листопада</month>

</dateStruct>

до редакції зателефонували з виробничого відділу видавництва «Донеччина».</> <s>Тоді <s><dateStruct value='12-11'>

<distanse reg='1 day' offset=after'11-11'>наступного дня</distanse>

</dateStruct>

до приміщення, де ще кілька днів тому перебувала редакція «Острова», прибули молоді люди бритоголової зовнішності та вражаючих габаритів.</>

У наступному прикладі використано елемент <exact>, щоб продемонструвати брак точності часового проміжку відносно дати:

<s><dateStruct >

<distanse exact =«N» offset=after> Після</distanse>

<year type='num'>1885 р.</year>

</dateStruct>

народники втратили свою провідну позицію у визвольному русі.</>

Подія, відносно якої називається анотована точка часу, може бути вираженя як датою (у попередніх прикладах), так і в інший спосіб, напр., *напередодні Нового року, після вітчизняної війни, за рік до помаранчевої революції* тощо. Для означення таких подій вживаємо елемент <occasion>:

<s>Національний рух не зник навіть
<dateStruct>
<distanse exact =«N» offset=after> після</distanse>
<occasion>придушення революції</occasion>
</dateStruct>
</>
<s> <dateStruct>
< distanse reg='1 day' offset=before> Напередодні</distanse>
<occasion>святвечора</occasion>
</dateStruct>
святий вогонь з Віфлеєма, який запалили від вічної лампади у печері, де народився Христос, привезли до Москви та Калінінграда.</>

Таким чином, у текстах корпусу виявляємо дати абсолютні та відносні, серед них точні й неточні, репрезентовані як вказівкою на століття, рік, місяць, день, так і вказівкою на подію (свято, історична епоха, часткова подія тощо). На початковому етапі кодування вважаємо за необхідне мінімально ідентифікувати дату в тексті, означивши її межі. Так, крім власне вказівки на точку часу (*1997, липень, XX, день народження* тощо), до дати належать елементи тексту *р., рік, ст., наприкінці, на початку, з середини, після, до* та ін.

Точні абсолютні дати:

<s>Прийняття <date>в грудні 1867 р.</date> конституції гарантувало (хоча б формально) рівність всіх мов та національностей.</>

<s>Як відомо, Крушельницький був автором низки публікацій у газеті «Індепендент», в яких наводилися уривки із допитів міліціонерів, які стежили за Георгієм Гонгадзе аж до дня його викрадення <date>16 вересня 2000 року</date>.</>

Неточні абсолютні дати, зокрема оказіональні:

<s>Так, англійській Ост-Індській та голандській Вест-Індській компаніям <date>на початку XVII ст.</date> держава надала виняткове (монопольне) право на торгівлю з Індією. </ >

<s><date>З самого початку 19 століття</date> український національний культурний рух був тісно пов'язаний з політичним.</ >

<s><date>Наприкінці XIX ст.</date> ринок чи не вперше за багатовікову історію свого існування та розвитку зіткнувся з серйозними проблемами. </ >

<s>Ідеї Братства, закладені у національну свідомість, отримали розвиток серед представників соціального руху <date>наприкінці 19 ст.</date>.</>

<s>В результаті розподілу селянських господарств <date>наприкінці 19 ст.</date> було створено ринок найманої праці.</>

<s><date>В середині 1850 рр.</date> селянський рух охопив 422 селища Київської, Катеринославської та Херсонської губерній.</>

<s><date>На святвечір</date> католики в різних регіонах Росії намагалися забути про кордони, що роз'єднують людей.</>

<s><date>Під час так званої "галилейської кризи"</date>, коли вчення Ісуса про живий хліб, який зійшов з неба, обурило навіть його наближених, і можливо тут і відбувся кардинальний перелом в душі Іуди. </>

<s>Відомий біблейський персонаж - Іуда Іскаріотський, якого ми знаємо як зрадника Ісуса Христа, став в <date>останні часи</date> об`єктом зацікавленості як вітчизняних так і зарубіжних дослідників Біблії та питань християнської релігії. </>

Неточність може або не може бути подолана залежно від характеру датованої події, напр., *країни підписали мирну угоду тільки в кінці століття* – точність дати може бути відновлена в струкутрованому поданні року і дня підписання угоди; *панівними монополістичні тенденції стають лише наприкінці XIX ст.* – подання обмежується приблизною вказівкою.

Абсолютні діапазонні дати (точні й неточні):

<s>Перші політичні організації з'явилися <date>в 1880–1890 рр.</date>.</>

<s>Взагалі монополістичні тенденції в різних формах та з різною силою проявлялися на всіх етапах розвитку ринкового суспільства (<date>з IV тисячоліття до н.е. до останньої третини XIX ст. </date>).</>

Відносні дати точні, зокрема оказіональні:

<s>Усього <date>через два роки після закінчення Великої Вітчизняної</date>.</>

Відносні дати неточні:

<s>Це сталося <date>за кілька днів до Різдва</date>.</>

Відносні дати діапазонні (точні й неточні):

<s>Зі шкільних часів ми пам'ятаємо історію про те, що експерти із капіталістичних країн після перемоги соціалістичної революції визначали наше відставання в 100 років.</> <s>Ми гордились тим, що наздогнали розвинені країни вже <date>через 10–20 років</date>.</>

Література

1. Демська-Кульчицька О. М. Основи Національного корпусу української мови. – К., 2005. – 219 с.

2. Електронний ресурс: http://nkum.nm.ru.
3. Електронний ресурс: http://www.w3.org/TR/1998/NOTE-datetime-19980827
4. Словник іншомовних слів / Уклад. Л. О. Пустовіт, О. І. Скопненко та ін. – К.: Довіра, 2000. – 1018 с.

ТЕГУВАННЯ ВЛАСНИХ НАЗВ ОРГАНІЗАЦІЙ В НАЦІОНАЛЬНОМУ КОРПУСІ УКРАЇНСЬКОЇ МОВИ

© Ніна Лозова, 2007
Інститут української мови НАН України (Київ)
УДК 811.161.2'374.72'22

У статті здійснено спробу уніфікувати оформлення власних назв у Національному корпусі української мови. Запропоновано способи тегування назв організацій.

В Інституті української мови продовжується робота над створенням Національного корпусу української мови. Однією з невідкладних проблем є визначення принципів тегування текстів. Завданням цієї статті є випрацювання способів тегування власних назв організацій.

Для того щоб виокремити в тексті власні назви організацій, потрібно визначити, що саме ми розуміємо під поняттям *організація*. Оскільки створюваний корпус української мови потребує формалізованого викладу даних, слід спиратися не на юридичне визначення, а на формальне. Пропонуємо власною назвою організації вважати будь-яку власну назву на позначення об'єднання осіб, організацій або держав. Це й підприємства, і громадські установи. До організацій зараховуємо й такі, як ЄС, СНД, Рада національної безпеки і оборони, Рада народних депутатів, Верховна Рада України, Рада Європи. Організацією вважаємо також комплекс споруд, у якому група людей або організацій здійснює певний обсяг робіт (це й аеропорт, і зоопарк).

Отже, назва організації – це назва об'єднання осіб, організацій або держав, перше слово якої пишеться з великої літери. Назва може бути оформлена графічно будь-яким чином, містити пунктуаційні знаки, скорочення. Усі слова (або частина слів), а не лише перше, можуть писатися з великої літери.

Слід визначити межу, за якою починається (закінчується) назва організації. Так, у наведених вище прикладах (*Верховна Рада України, Рада Європи*) до назви організації належить і географічна назва. Однак якщо топонім лише вказує на місце, якому організація належить чи де вона