КОМП'ЮТЕРНА ЛІНГВІСТИКА

ПРИНЦИП КОДУВАННЯ АБРЕВІАТУР У НАЦІОНАЛЬНОМУ КОРПУСІ УКРАЇНСЬКОЇ МОВИ

© Валентина Балог, 2008
к. філол. н., Інститут української мови НАН України (Київ)
УДК 161.2.81'72'22

У статті проаналізовано й описано принципи кодування засобами TEI (text encoding initiative) скорочених назв української мови у корпусі текстів.

Абревіатура – поширений елемент будь-якого тексту, який уживається більшою чи меншою мірою практично в усіх мовних стилях. Явище абревіації в мові має глибоке коріння, спричинене потребою економії засобів, часу, зусиль та використовуване як у писемному (первинно), так і в усному мовленні.

Абревіатура (від лат. abbrevio – скорочую) – складноскорочене слово, утворене з початкових складів або з перших літер слів словосполучення [4; 9].

Предметом нашого дослідження є формалізація абревіатур у текстах, відібраних для Національного корпусу української мови, на основі вже розроблених в лінгвоукраїністиці класифікаційних та структурних характеристик скорочених назв. Вона зумовлена потребою адаптованого та уніфікованого кодування абревіатур, використовуваних в українській мові, у підготовленні лінгвістичного матеріалу для корпусу текстів для чіткого розуміння їхньої будови та функції комп'ютером та, як наслідок, адекватної відповіді на можливі запити користувача. Мета дослідження – виробити чітку та уніфіковану методику кодування абревіатур. З огляду на поставлене завдання, ми обмежимося тільки загальною лексикологічною характеристикою скорочених слів.

Явище абревіації добре досліджене в українській мові: його розглядають у загальних лінгвістичних працях, вивчають в окремих спеціальних дослідженнях. Поняття абревіатури витлумачено в загальних та галузевих словниках. В українському мовознавстві наявні різні класифікаційні характеристики скорочених назв за формальними та функціональними ознаками [5; 6; 7; 9], основу яких складають три основні типи: складові, ініціальні та змішані слова.

Найпоширенішими в мові є лише два види абревіатур: 1. Слова, утворені з початкових літер та звуків, що входять до складу багаточленного найменування, наприклад: *ООН* (Організація Об'єднаних Націй), *ВАК* (Вища атестаційна комісія) тощо; їхня назва – акроніми. Часто такі скорочені назви свідомо прагнуть уподібнити за милозвучністю до звичних нескорочених слів мови, інколи ігноруючи навіть повну омонімію, наприклад: *АМУР* (автоматична машина управління і регулювання), *КАСКАД* (клас автоматизованих систем комплексного аналізу документів), *МАРС* (машина автоматичної реєстрації і сигналізації), *бор* (батарея оптичної розвідки). Ця обставина спричинює належність їх до активного словника сучасного мовлення, переважно в науковій термінології. 2. Абревіатури, утворені складанням частин слів, наприклад: *райвно, технагляд, автобаза* тощо. Широко використовуються в мові спеціальні абревіатури, поєднані з цифрами: *АН-70, АН-124* тощо. Останні складають переважно технічну номенклатуру – назви серійних моделей різноманітних апаратів і приладів.

Згідно зі стандартами мови електронного розмічування SGML, абревіатуру, незалежно від типологічних характеристик, за схемою TEI передбачено кодувати як коротко, тобто виділяти тільки саму абревіатуру, так і за принципом повної інформації, тобто подавати функціональне (семантичне навантаження), формальне (тип скорочення) та пояснювальне (розшифрування) значення абревіатури. Усю необхідну інформацію подаємо в межах елемента <abbr>, який за Принципами TEI маркує довільне скорочення. У разі розширеного тегування повна інформація надається через атрибути: пояснювальний ***expan*** і формально-функціональний ***type***. Глобальний атрибут ***type*** детермінує тип та функцію скорочення відповідно до прийнятої TEI класифікації із значеннями: ***contraction*** – стягнена форма, ***suspension*** – пропуск, три крапки, ***superscription*** – верхній індекс, ***brevigraph*** – скорочений запис і ***acronym*** – акронім; також значення ***title*** – назва в адресі, ***geographic*** – географічна назва, ***organization***[1] – назва організації тощо, які можуть бути поєднані для одного слова [1; 164], наприклад: <abbr type=org type=acronym expan='Національний авіаційний університет'>НАУ</abbr>. Атрибут ***expan*** використовується для розшифрування абревіатури. Вважаємо за доцільне застосовувати цей засіб розширення інформаційного поля, зважаючи на дослідницький тип корпусу текстів. Наприклад: <abbr type=org expan='районний відділ народної освіти'> райвно</abbr> [10].

[1] Перелік значень атрибутів може бути продовжений упорядниками залежно від очевидних потреб.

Визначення методу кодування абревіатур і взагалі скорочених варіантів слів тісно переплітається з питаннями кодування інших груп лексики, як-от власні назви, часові та числові показники, оскільки кодуванню підлягає скорочена назва в контексті. Саме ця обставина формує проблему корпусної адаптації текстів, а саме проблему поєднання різних кодів у тих сегментах тексту, де поєднуються різні групи лексики. Щодо кодування, наприклад, дат [8; 82-87], де можуть використовуватися скорочення на зразок р. (рік), рр. (роки), ст. (століття), год. (година), с. (секунда) і под., вважаємо прийнятним не виокремлювати ці об'єкти. На нашу думку, детальне кодування будь-яких скорочень призведе до надлишку метаінформації, перевантаження метатекстової частини корпусу і супроводжуватиметься надмірними зусиллями упорядників корпусу. Тому в межах проекту НКУМ пропонуємо обмежитися кодуванням абревіатур, які подають та розшифровують у загальногалузевих та загальних словниках, а також коли є потреба виділення такої назви.

Важливим питанням є визначення методики кодування власних назв, зокрема організацій [3; 88-92], до складу яких часто входять абревіатури, на відміну від назв – повних абревіатур, кодування яких не викликає сумнівів. У кодуванні власних назв з використанням абревіації можна йти двома рівноадекватними шляхами: 1) виділяти назву як ім'я, використовуючи елемент ***name*** та відповідні атрибути ***type***, ***abbr*** з різними значеннями для уточнення форми подання назви та власне поняттєвого навантаження, наприклад: <name type=org type=abbr expan='Українська православна церква Московського патріархату'> УПЦ МП</name>; 2) виділяти назву як скорочення, застосовуючи в такому разі методику кодування абревіатур. І в першому, і в другому випадках буде зрозумілим виділення скороченого слова. Допускаємо, що в процесі верифікації корпусу постане необхідність повної уніфікації тегування однотипних елементів тексту, зокрема абревіатур, однак на теоретичному рівні різні способи кодування є правильними.

За наявності абревіатури в будь-якій назві вважаємо за доцільне окремо її тегувати, проте обмежитися тільки пояснювальною характеристикою скороченої назви за допомогою атрибута ***expan***, не подаючи при цьому ні формального, ні функціонального уточнення, оскільки воно вже закладене кодуванням самої назви, напр.:

<s> Голова правління акціонерного товариства Дмитро Кадельник проінформував їх, а також представників обласної та міської влади, <name type=org>Фонду <abbr expan='державного майна'>Держмайна</abbr></org>, науковців, журналістів, що реструктуризацію підприємства планується здійснювати у три етапи: перший – проведення організаційної, виробничої та

фінансової санації підприємства, другий – створення умов для розвитку високопродуктивного та ефективного індустріального парку; організація 100-150 нових підприємств на 100 тисячах квадратних метрів виробничих площ; досягнення ними через 3-4 роки щорічного рівня продажу у межах 150 мільйонів гривень. </s>.

Взагалі в процесі кодування корпусу на лексичному рівні рішення про пріоритети елементів TEI варто приймати окремо для кожного конкретного випадку відповідно до специфіки тексту та його стильового навантаження, оперуючи адаптованими до української мови прийомами тегування.

Література

1. Демська-Кульчицька О.М. Основи національного корпусу української мови. – К., 2005. – 219 с.
2. Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна та ін.. – К.: Довіра, 2005. – 471 с.
3. Лозова Н.Є. Тегування власних назв організацій в Національному корпусі української мови //Лексикографічний бюлетень: Зб. наук. праць. – К, 2007. – Вип. 16. – С. 88–91.
4. Словник іншомовних слів. / За. ред. О.С. Мельничука. – К.: Гол. ред. УРЕ, 1977. – 776 с.
5. Сучасна українська літературна мова. / За ред. І. К. Білодіда. – Кн.: 5. Лексика і фразеологія. – К.: Наук. думка, 1973. – 439 с.
6. Сучасна українська літературна мова / За ред. М. Плющ. – К.: Вища шк., 2001.
7. Сучасна українська мова / За ред. О. Д. Пономарева, – К.: Либідь, 2001. – 400 с.
8. Тищенко О. М. Тегування дат в Національному корпусі української мови//Лексикографічний бюлетень: Зб. наук. праць. – К, 2007. – Вип. 16. – С. 82–87.
9. Українська мова. Енциклопедія. – К.: Укр. енциклопедія, 2000. – 750 с.
10. http://www.tei.org

КОРПУСИ ЖЕСТОВИХ МОВ ГЛУХИХ: СВІТОВИЙ ДОСВІД

© Оксана Тищенко, 2008
к. філол. н., Інститут української мови НАН України (Київ)
УДК 161.2.81'322.221.24+ 72'22

У статті здійснено огляд інтернет-ресурсів, присвячених корпусам жестових мов. Виявлено основні мотиви й цілі створення таких анотованих баз даних у США, Нідерландах, Німеччині, Японії та Греції, проаналізовано основні принципи добору й анотування відеоінформації.

Автоматичне розпізнавання й машинне статистичне опрацювання жестової мови глухих, створення відповідних систем оброблення даних –