

Morphological Tagging of Verbforms at Corpora of English Texts

Kotyk Serhii

Laboratory of Computational Linguistics, Kyiv National
Linguistic University, UKRAINE,
Kyiv, Velyka Vasylkivska street 73,
E-mail: kotykserg@gmail.com

The present article is about morphological tagging of English verbforms based on the deductive calculus of the English conjugation paradigm.

The first part is devoted to relevance of the topic explaining the importance of the topic and reasoning the choice of this topic. This part is followed by some general theoretical notes including some information about the verbforms classification.

In 1986–1991 a large research team under the guidance of the Laboratory of statistical lexicography of the Kyiv state institute for foreign languages (prof. V. Perebeynoss) carried out a large-scale statistical investigation of English conjugation forms in four text corpora: fiction (F), drama (D), scientific texts (S) and periodical political press (P).

The deductive calculus of the English conjugation paradigm on the basis of 14 differential features resulted in establishing 526 forms.

The main part is about the application that shows the practical use of current research. This part includes general knowledge of how the application works including stages of application's work with explanation of what exactly happens in each stage so that a reader could understand the basic principles. There are also screenshots that illustrate the application interface.

In written English predicate is a principal part of a sentence so there is hardly a sentence without it. This combined with existence of non-finite verbforms makes the verb one of or even the most frequent part of a sentence. That is why none of morphological analyzer can avoid classifying and tagging verbforms.

Current research illustrates the possibility of automatic verbform recognition in random English text while the existence of the software described in this article proves the possibility of comparatively accurate formalization of English verbforms search and recognition.

Морфологічне кодування дієслівних форм в корпусі англійських текстів

Котик Сергій

Лабораторія комп'ютерної лінгвістики, Київський
національний лінгвістичний університет,
УКРАЇНА, м.Київ, вул.Велика Васильківська, 73,
E-mail: kotykserg@gmail.com

В статті подані загальні засади, прийняті за основу для розробки автоматичної системи кодування граматичних форм англійського дієслова. Розроблене програмне забезпечення ґрунтується на 14 формальних диференційних ознаках та базі даних граматичних форм англійських дієслів. Створена база даних нараховує понад 800 дієслів. Описано роботу програми морфологічного кодування англійських дієслів включаючи основні принципи та етапи. Розроблена програма здійснює автоматичний пошук дієслівних форм у нерозміченому тексті і приписує відповідний граматичний код.

Ключові слова – кодування, теґування, код, дієслівна форма, програмне забезпечення.

I. Вступ

Як відомо, одним із найважливіших етапів засвоєння системи англійської мови є вивчення дієслів. У мовній системі англійської мови дієслово займає особливе місце, не властиве жодній іншій частині мови. Усі англійські висловлювання формуються на базі дієслівного члена, що і представляє описувану ситуацію.

Актуальність даної роботи полягає в тому, що проблеми лінгвістичних досліджень, пов'язаних із дієсловом, мають давню традицію і привертають увагу дослідників від античності до сучасного часу. Зокрема на сучасному етапі, під час автоматичного опрацювання англійських текстів, уникнути проблеми розмежування та ідентифікації дієслівних форм неможливо. Отже, автоматичне кодування дієслівних форм має дуже широкий спектр застосування і є надзвичайно корисним інструментом в руках лінгвіста.

II. Морфологічне кодування дієслівних форм

Кодування частин мови, або теґування – це процес встановлення та приписування кожному слову, або окремій словоформі відповідного граматичного класу. Якщо раніше цей процес здійснювався вручну, то зараз в більшості випадків цим займаються машини.

Історія мовознавства знає немало різних класифікацій англійської дієслівної лексики. Можна впевнено стверджувати, що однозначних класифікацій не існує і навряд чи вони можуть бути розроблені.

У даній роботі за основу класифікації було прийнято схему розроблену раніше під керівництвом професора В.С. Перебийніс [1]. У 1986–1991 рр. було здійснено статистичне обчислення системи форм

англійського дієслова на матеріалі текстів загальним обсягом 12 млн. слововживань: художньої прози (F), драми (D), наукових текстів (S) і публіцистики (P) [2]. Дана класифікація базується на 14 диференційних формальних ознаках і налічує 526 основних інваріантних кодів, кожному з яких відповідає певна аналітична форма англійського дієслова.

III. Розробка програмного забезпечення мовою C#

Як вже зазначалося вище, тема даної роботи має досить багато практичних застосувань. Одним з цих застосувань є здійснення автоматичного пошуку, розмежування і кодування форм англійського дієслова в нерозміченому тексті. Зазначений процес потребує великих зусиль аналізатора та багатостороннього і багатоетапного опрацювання тексту.

Розробка програмного забезпечення орієнтувалася на здійснення пошуку дієслівних форм в довільному англійському тексті. Тобто, при знаходженні певної дієслівної форми в тексті та її ідентифікації, програма мала поруч з головним словом, тобто лексичною частиною аналітичної форми, приписати відповідний код цієї форми.

Роботу програми можна умовно розділити на декілька етапів:

1. Здійснення сегментації тексту на частини. На цьому етапі програма автоматично сегментує текст на частини спираючись на список сполучників типу: “and”, “also”, “since” та ін. та розділових знаків: “:”, “.”, “,” та ін., що занесені до окремої бази даних програми.

2. Формальний аналіз частин тексту. На етапі формального аналізу довільного тексту програма аналізує кожен сегмент тексту, як наприклад: “...he had left home earlier ...” і приписує кожній складовій дієслівної форми і дієслівній формі в цілому відповідні коди, які занесені до бази даних програми, наприклад: “... I had<aux> left<verb:Participle2<163>> home earlier ...”. Такий запис означає, що “had” -- допоміжне слово, “left” – головне слово, яке виражене дієсловом у формі минулого часу, а саме Participle II. Код “163” містить в собі інформацію про всю дієслівну форму разом, в данному випадку, що таке поєднання допоміжного дієслова та головного відповідає аналітичній дієслівній формі під кодом 163 [1].

3. Виведення результатів у вигляді повноцінного тексту. Виведення результатів морфологічного аналізу дієслівних форм є ніщо інше як зклеювання всіх сегментів до купи з урахуванням внесених змін та виділення кольором важливих частин (див. Рис. 1).

4. Статистичне опрацювання результатів. Етап підрахування результатів є підготовчою стадією для наступного етапу. На цьому кроці програма підраховує частоту вживання кожної дієслівної форми на масиві всього аналізованого тексту.

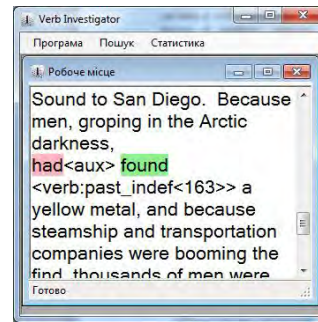


Рис. 1 “Приклад виведення результатів ”

5. Виведення статистичних результатів.

Статистичні результати програма виводить у вигляді таблиці. У лівий стовпчик занесені коди, а у правий їх частота вживання в аналізованому тексті (див. Рис. 2).

Код	Частота
63	56
251	18
45	16
88	8
163	7
263	4
139	3
4	3
127	3
61	2

Рис. 2 "Вікно виводу статичних результатів"

Висновок

У письмовому англійському тексті присудок є обов’язковим членом речення, а якщо врахувати такі дієслівні форми, як Participle 1, Participle 2 та Gerund, то дієслово має високі показники частоти. Саме тому жодна система морфологічного аналізу не може оминати питання розмежування та тегування дієслівних форм. Дане дослідження демонструє можливість автоматичного розпізнавання дієслівних форм у довільному англійському тексті. Реалізація описаної вище програми доводить можливість відносно якісної формалізації пошуку та розпізнавання форм англійського дієслова.

Література

- [1] English Conjugation: System And Functioning. Reference-book. /Ed. Perbyiniss V. – Russian State University for the Humanities. – Moscow, 2008. – 545 p.
- [2] Perbyiniss V., Khidekel S. Morphology of the English Verb: System and Functioning – A Lexico-Grammatical Reference Book/ New Trends in Lexicography: Ways of Registrating and Describing Lexis. Part II, Chapter XVII. – Ed. O. Karpova, F. Kartashkova. – Cambridge Scholars Publishing, 2010. – P. 191-203.