УДК 004.912 doi: 10.20998/2227-6890.2017.52.20

С. В. ПЕТРАСОВА, М. О. КУЗЬМІНА, І. О. МАНУЙЛОВ

ОСОБЛИВОСТІ МОРФОЛОГІЧНОЇ РОЗМІТКИ КОРПУСІВ УКРАЇНСЬКОЇ МОВИ НА ПРИКЛАДІ ТЕХНІЧНОЇ ДОКУМЕНТАЦІЇ

У статті розглядаються особливості автоматичної морфологічної розмітки корпусів текстів української мови. Створено корпус текстів української мови, які представляють інструкції технічної документації. Розроблено систему тегів для формалізації морфологічної інформації. Описано розроблену програмну реалізацію запропонованого методу автоматичної морфологічної розмітки, що дозволяє отримувати з корпусу технічної документації приклади вживання в мові як конкретних словоформ, так і слів у всіх їх граматичних формах.

Ключові слова: морфологічна розмітка, корпус текстів, технічна документація, тегсет.

В статье рассматриваются особенности автоматической морфологической разметки корпусов текстов украинского языка. Создан корпус текстов украинского языка, представляющих инструкции технической документации. Разработано систему тегов для формализации морфологической информации. Описано разработанную программную реализацию предложенного метода автоматической морфологической разметки, позволяющую получать из корпуса технической документации примеры употребления как конкретных словоформ, так и слов во всех их грамматических формах.

Ключевые слова: морфологическая разметка, корпус текстов, техническая документация, тегсет.

The article deals with the features of tagging Ukrainian corpora. Representing instructions for technical documentation, the corpus of Ukrainian texts is created. The main principles and means of morphology in the Ukrainian language are determined for further tagging. The system of tags is developed to formalize morphological information. Based on the procedural method of morphological analysis, the algorithm for tagging nouns in the Ukrainian language is described. The implementation of the proposed method of automatic tagging is developed. This result allows receiving examples of the use of both specific word forms and words in all their grammatical forms from the technical documentation corpus.

Keywords: tagging, text corpus, technical documentation, tagset.

Вступ. В останні десятиліття все більш активно впроваджуються методи дослідження, що базуються на корпусах текстів. В сучасній лінгвістиці під корпусом розуміють обмежений за обсягом набір електронних текстів, зібраних з метою максимально точного представлення досліджуваного варіанта мови [1]. Використання корпусів дозволяє вивчати одиниці тексту, слугує джерелом і інструментом багатоаспектних лексикографічних праць та джерелом для уточнення існуючих граматик і складання нових.

Слід зазначити, що основна особливість мовних корпусів – «розміченість», тобто наявність у складі текстів спеціальних міток, що описують як самі тексти, так і одиниці, що відносяться до різних мовних рівнів. Під час автоматичної обробки природномовних текстів саме морфологічна розмітка є основою як для морфологічного аналізу, так і для подальших форм аналізу – синтаксичного і семантичного.

Аналіз останніх досліджень і публікацій. В процесі розвитку мережі Інтернет стали доступні великі обсяги текстового матеріалу, придатного для проведення різних лінгвістичних досліджень. При цьому постає питання щодо репрезентативності і збалансованості мовного матеріалу, який є ключовим при формуванні корпусів текстів [2].

Серед сучасних корпусів української мови варто зазначити національний корпус української мови Інституту української мови НАНУ (50 млн. слововживань) – вибірка текстів сучасної української мови, яка є репрезентативною для всіх функціональних рівнів загальнонародної мови та призначена для лінгвістичного аналізу й технологічного застосування [3].

Корпус сучасної української мови обсягом 3 млн. словоформ, побудованого інститутом філології Київського національного університету ім. Тараса Шевченка, надає інформацію щодо конкордансів, за допомогою яких можна вивчати особливості використання слів у текстах різних стилів, асоціативні зв'язки між словами, кількісні характеристики вживання у текстах мовних одиниць, що розкривають закономірності лексичної та статистичної будови текстів, функціонування мови в мовленні, стилістичні та граматичні особливості [4].

Корпус текстів з комп'ютерної лінгвістики лабораторії комп'ютерної лінгвістики Київського національного лінгвістичного університету, обсягом понад 4 тис. слововживань, містить стандартний набір програм морфологічного кодування, в якому здійснюється пошук за словоформою та граматичним кодом [5].

До загальнодоступних корпусів української мови також входять: лексикографічна система «Український національний лінгвістичний корпус» Українського мовно-інформаційного фонду НАН України [6], корпус українських текстів ДонНУ [7], Браунський корпус української мови СЄНУ імені Лесі Українки [8] та ін.

Найважливішим складником корпусів є розмітка. Залежно від цілей дослідження, автоматична обробка корпусів текстів може включати як глибоку синтаксичну та семантичну розмітку, так і обмежуватися лише морфологічною розміткою. Здійснення морфологічної розмітки корпусних текстових даних попередньо передбачає: побудову тегів, які через формальний запис експлікують граматичні значення слів, до яких вони приписані; створення тегсету з відповідною семантикою, засобами якого адекватно детерміновано для кожної лексичної одиниці тексту її відношення до морфологічної системи мови [9].

Дослідження показало, що існують чотири основні критерії побудови тегів:

1) довжина: короткі символи зручніші для аналізу, ніж довгі;

2) експліцитність: символи, які легко інтерпретувати і розуміти, є зручнішими для використання;

3) аналітичність: символи, які підлягають декомпозиції на логічні складники, кращі, ніж ті, які не можна декомпонувати, наприклад, тег *NP*1 може бути розкладений на *N* = іменник, *P* = власна назва, 1 = однина. Дотримання критерію аналітичності дозволяє здійснювати корпусні дослідження навіть за умови різного рівня їхньої деталізації. Так, символом *N** можна задати усі іменники, і далі, деталізуючи: *N**1 – усі іменники в однині, *NP** – усі іменники власні назви і т.д.;

4) однозначність: у межах тега унікальні символи співвідносяться з унікальними значеннями, наприклад, *N* – іменник, *A* – прикметник, *P* – займенник, на першій позиції у коді неможна використовувати позначення якихось інших значень у цій позиції.

Морфологічні теги є передусім засобом формалізації морфологічної інформації і призначені саме для програмного оброблення.

Серед сучасних систем морфологічної обробки найбільш відомими є:

– «Mystem», яка працює на основі словника і здатна формувати морфологічні гіпотези про незнайомі слова [10];

– «Stemka», яка у своїй роботі використовує імовірнісний підхід [11];

– «Морфер» виконує відмінювання російських та українських словосполучень за відмінками, визначення статі за прізвищем, ім'ям, відмінювання чисел [12];

– «ОРФО», яка виконує пошук однієї форми слова за іншою його формою [13].

Незважаючи на значну кількість програмних засобів, їх функціональність, більшість систем не адоптовані для української мови, у зв'язку з чим є актуальним створення системи автоматичного опрацювання української мови для проведення аналізу в першу чергу морфологічних показників мови.

Метою цього дослідження є розробка методу морфологічної розмітки корпусів технічної документації української мови. Використання цього методу дозволяє автоматизувати обробку природномовної інформації для подальшого аналізу та використання корпусу.

Матеріали і результати дослідження. Основні завдання укладання корпусу текстів передбачають формулювання лінгвістичної концепції корпусу, визначення предметної галузі та парадигми даних корпусу, проектування корпусу, визначення параметрів анотування даних і лінгвістичне забезпечення програмної обробки.

В результаті проведеного огляду методів і підходів автоматичного морфологічного аналізу корпусів української мови було розроблено алгоритм, в основі якого лежить безсловниковий процедурний метод морфологічного аналізу [14].

Метод морфологічного аналізу текстів використовує таблиці суфіксів, закінчень та список службових незмінних слів – прийменників. Запропонований метод характеризується високою швидкістю визначення словоформ за рахунок використання словника готових закінчень.

Реалізація морфологічної розмітки корпусу української мови на прикладі технічної документації полягає в здійсненні послідовності наступних етапів.

На першому етапі здійснюється відбір джерел текстів для створення корпусу. Згідно зі стандартами коректної побудови створено корпус україномовних текстів, які представляють інструкції технічної документації. Створений корпус володіє такими ознаками як: репрезентативність, збалансованість, відібраність, машиночитаність та стандартність.

На другому етапі розроблено систему тегів для символічного позначення частини мови та морфологічних ознак у корпусі технічної документації.

У морфологічну структуру української мови входять парадигми відмінюваних частин мови та морфемна структура усіх частин мови, тобто всіх класів слів, як змінних, так і незмінних. Кожна відмінювана частина мови має свої характерні парадигматичні мікросистеми, які створюють систему парадигм даної частини мови, що в своїй сукупності складають загальну морфологічну парадигматичну систему української мови. Саме в частинах мови найпоказовіше відображаються особливості морфологічного ладу української мови, зокрема сукупність морфологічних категорій та їхніх грамем, словозмінна морфеміка, співвідношення синтетизму й аналітизму в морфологічній структурі мови.

На основі розглянутої морфологічної парадигматичної системи української мови для здійснення морфологічної розмітки було обрано іменник як центральну частину мови в українській мові.

При побудові тегів були враховані критерії довжини, експліцитності, аналітичності та однозначності.

Наступним кроком є побудова лінгвістичної бази даних, яка складається з таблиць морфологічних категорій, можливих суфіксів і закінчень іменників, та прийменників.

На останньому етапі здійснюється автоматичне виділення токенів та розмітка корпусу текстів:

– після виділення в словах флексій для кожного слова знаходиться відповідність у таблиці суфіксів та закінчень;

– у разі відповідності слову приписується частина мови та морфологічні ознаки у вигляді тегсету.

Наприклад, результат автоматичної розмітки тексту інструкції для слова «налаштування» у вигляді тегсету:

<*pos*="*N*" *gram*="I, *n*, *s*, *f*|*m*|*c*" | "II, *n*|*g*|*a*|*v*, *s*|*p*, *m*|*n*" |
| "IV, *n*|*g*|*a*|*v*, *s*, *n*" />,

де *pos* – частина мови; *gram* – граматичні категорії; *N* – іменник; I – 1 відміна іменників; II – 2 відміна іменників; IV – 4 відміна іменників; *n* – називний відмінок; *g* – родовий відмінок; *a*– знахідний відмінок; *v* – кличний відмінок; *s* – однина; *p* – множина; *f* – жіночий рід; *m* – чоловічий рід; *c* – середній рід.

Всі наведені етапи автоматичної морфологічної розмітки корпусу текстів реалізовано у вигляді прикладної програми (рис. 1).

Розроблена система автоматичної морфологічної розмітки дозволяє отримувати з корпусу технічної документації приклади вживання в мові як конкретних словоформ, так і слів у всіх їх граматичних формах.

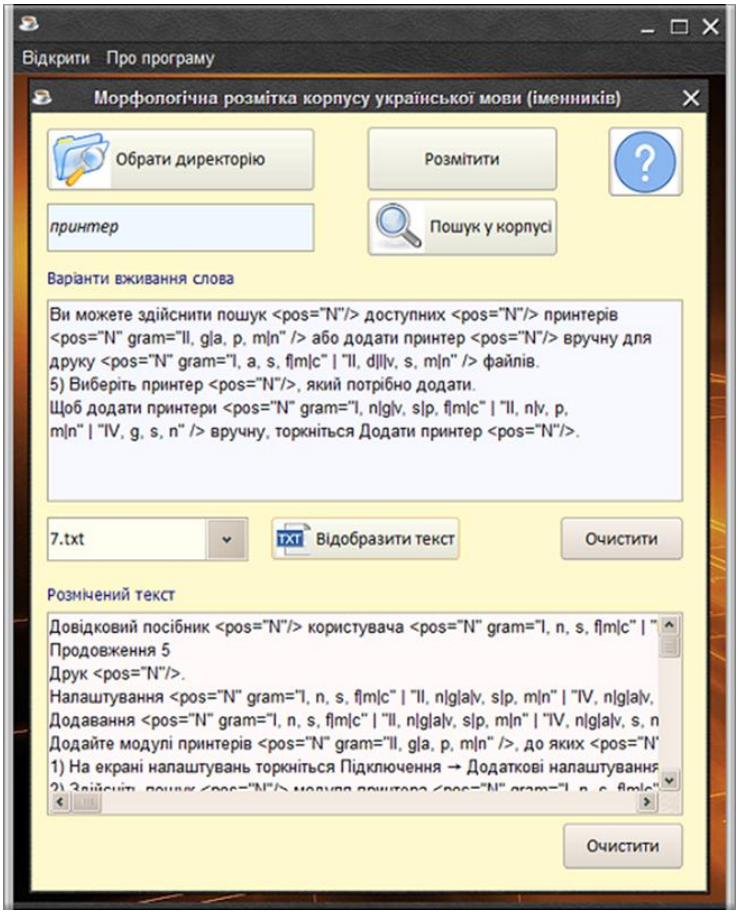

Рисунок 1 – Прикладна програма автоматичної морфологічної розмітки корпусу технічної документації української мови

Висновки. Дане дослідження дозволяє наглядно побачити реалізацію завдання створення власного корпусу. Згідно з інформацією про корпуси та технічну літературу виведено параметри, яких необхідно дотримуватися, щоб корпус був найбільш актуальним і відображав зміст технічної документації найдостовірніше.

В роботі розглянуто метод морфологічної розмітки корпусів, застосування якого дозволяє підвищити якість розмітки масивів україномовних текстів, що містять велику кількість слів.

В результаті дослідження розроблено програмне забезпечення, яке дозволило уникнути зайвих помилок за рахунок конкретизованої системи кодування та зменшило неоднозначність на морфологічному рівні аналізу тексту. Таким чином, встановлення певної типології метаданих сприяло запобіганню дослідницьких непорозумінь у процесі зіставлення та опрацювання корпусів різних текстів.

Реалізація найважливіших параметрів морфологічних тегів – аналітичності та однозначності, що дозволила провести лінгвістичне анотування, або розмітку, може слугувати базою для подальших лінгвістичних досліджень у корпусній лінгвістиці. Зокрема, на основі розміченого корпусу можна отримати дані про частоту лексем, словоформ, граматичних категорій, прослідкувати зміну частот і контекстів в різні періоди часу, отримати дані про спільну зустрічальність лексичних одиниць і т.п.

Список літератури: **1.** *Герд А.С.* Прикладная лингвистика / *А.С. Герд* – СПб. : Изд-во С.-Петерб. ун-та, 2005. – 268 с. **2.** *Большакова Е.И.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / *Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ и др.*– М. : МИЭМ, 2011. – 272 с. **3.** *Демська-Кульчицька О.М.* Основи національного корпусу української мови / *О.М. Демська-Кульчицька.* – К. : Інститут української мови НАНУ, 2005. – 219 с. **4.** Корпус текстів української мови [Електронний ресурс]. – Режим доступу : http://www.mova.info/corpus.aspx?l1=209. – Дата звертання : 25 жовтня 2017. **5.** *Бобкова Т.В.* Корпус текстів з комп'ютерної лінгвістики / *Т.В. Бобкова та ін.* // Комп'ютерні науки та інформаційні технології : матеріали 4-ї Міжнародної науково-технічної конференції, 17 жовтня 2009 р. – Львів, 2009. – С. 405–407. **6.** Український національний лінгвістичний корпус [Електронний ресурс]. – Режим доступу : http://unlc.icybcluster.org.ua/virt_unlc/ – Дата звертання : 25 жовтня 2017. **7.** *Данилюк І.Г.* Корпус текстів для вивчення граматичної службовості: класифікація граматичних класів і підкласів / *І.Г. Данилюк* // Лінгвістичні студії. – Донецьк : ДонНУ, 2013. – № 27. – С. 221–229. **8.** *Старко В.Ф.* Формування браунського корпусу української мови / *В.Ф. Старко* // Мовні і концептуальні картини світу. – 2014. – № 48. – С. 415–421. **9.** *Бабина О.И.* Автоматизация лингвистической разметки корпуса текстов [Електронний ресурс] / *О.И. Бабина, Н.Ю. Дюмин.* – Режим доступу : http://helling100./pubs/AutomationBabinaDyumin.pdf. – Дата звертання : 25 жовтня 2017. **10.** Система Mystem [Електронний ресурс]. – Режим доступу : https://tech.yandex. ru/mystem/ – Дата звертання : 25 жовтня 2017. **11.** Система Stemka [Електронний ресурс]. – Режим доступу : http://linguist.nm.ru/stemka/ stemka.html – Дата звертання : 25 жовтня 2017. **12.** Програма відмінювання [Електронний ресурс]. – Режим доступу : http://morpher.ru/ DemoUA.aspx. – Дата звертання : 25 жовтня 2017. **13.** Система ОРФО – [Електронний ресурс]. – Режим доступу : http://www.orfo. ru/features/ – Дата звертання : 25 жовтня 2017. **14.** *Бабина О.И.* Корпусный метод автоматического морфологического анализа флективных языков / *О.И. Бабина, Н.Ю. Дюмин* // Вестник Южно-Уральского гос. ун-та, 2012. – № 25. – С. 38–44.

References (transliterated): 1. Gerd A.S. *Prikladnaya linguistika* [Applied linguistics]. SpB, SpB University, 2005. 268 p. **2.** Bolshakova E.I. *Automaticheskaya obrabotka tekstov na yestestvennom yazyke i komputernaya lingvistika* [Automatic processing of texts in natural language and computational linguistics]. MIEM, 2013. 272 p. **3.** Demska-Kulchytska O.M. *Osnovy nationalnogo korpusu ukrainskoi movy* [Fundamentals of the National Corpus of the Ukrainian Language]. Instytut ukrainskoi movy NANU, 2005. 219 p. **4.** Korpus textiv ukrainskoi movy [The corpus of texts of the Ukrainian language]. Available at: http://www.mova.info/corpus.aspx?l1=209. (accessed 25.10.2017). **5.** Bobkova T.V Korpus tekstiv z kompyuternoii lingvistiki [The corpus of texts on computational linguistics]. *Komputerni nauky ta informatsiyni technologii: materialy 4 Mizhnarodnoi naukovo-technichnoi konferencii* [Computer Science and Information Technologies: Materials of the 4th International Scientific and Technical Conference]. Lviv, 17.10.2009. pp. 405–407. **6.** *Ukrainskiy nationalniy lingustychny korpus* [Ukrainian national Linguistic Corpus]. Avaliable at : http://unlc.icybcluster.org.ua/virt_unlc/. (accessed 25.10.2017). **7.** Danylyuk I.G. Korpus tekstiv dlya vyvchennya gramatychnoi sluzhbovosti: klasyfikatsia gramatychnykh klasiv i pidklasiv [The corpus of texts for the study of grammar: the classification of grammatical classes and subclasses]. *Linguistychny studii*. Donetsk, DonNU, 2013, no. 27, pp. 221–229. **8.** Starko V.F. Formuvannya braunskogo korpusu ukrainskoi movy [Formation of the Brown corpus of the Ukrainian language]. *Movni i kontseptualni kartyny svitu*. 2014, no. 48, pp. 415–421. **9.** Babina O.I. *Avotatizatsya linguisticheskoy razmetki korpusa tekstov* [Automation of linguistic tagging of the corpus of texts]. Available at : http://helling100./pubs/AutomationBabinaDyumin.pdf. (accessed 25.10.2017). **10.** *Systema Mystem* [Mystem system]. Available at: https://tech.yandex.ru/mystem/. (accessed 25.10.2017). **11.** *Systema Stemka* [Stemka system]. Available at: http://linguist.nm.ru/stemka/stemka.html. (accessed 25.10.2017). **12.** *Programa vidminyuvannia* [Declination program]. Available at: http://morpher.ru/DemoUA.aspx. (accessed 25.10.2017). **13.** *System ORFO*]ORFO system]. Available at: http://www.orfo.ru/features/. (accessed 25.10.2017). **14.** Babina O.I. Korpusny metod autamticheskogo morfologicheskogo analiza flektivnych yazykov [Corpus method of automatic morphological analysis of inflexional languages]. *Vestnik Yuzhno-Uralskogo gos. Universiteta*. 2012, no. 25, pp. 38–44.

Надійшла (received) 15.12.2017

Бібліографічні описи / Библиографические описания / Bibliographic descriptions

Особливості морфологічної розмітки корпусів української мови на прикладі технічної документації / С. В. Петрасова, М. О. Кузьміна, І. О. Мануйлов // Вісник НТУ «ХПІ». Серія : Актуальні проблеми розвитку українського суспільства. – Харків : НТУ «ХПІ», 2017. – № 52 (1273). – С. 114–117. – Бібліогр. : 14 назв. – ISSN 2227-6890.

Особенности морфологической разметки корпусов украинского языка на примере технической документации / С. В. Петрасова, М. А. Кузьмина, И. А. Мануйлов // Вісник НТУ «ХПІ». Серія : Актуальні проблеми розвитку українського суспільства. – Харків : НТУ «ХПІ», 2017. – № 52 (1273). – С. 114–117. – Бібліогр. : 14 назв. – ISSN 2227-6890.

Features of tagging Ukrainian corpora by the example of technical documentation / S. V. Petrasova, M. O. Kuzmina, I. O. Manuilov // Bulletin of NTU "KhPI". Series : Actual problems of Ukrainian society development. – Kharkiv : NTU "KhPI", 2017. – No. 52 (1273). – P. 114–117. – Bibliogr. : 14. – ISSN 2227-6890.

Відомості про авторів / Сведения об авторах / About the Authors

Петрасова Світлана Валентинівна – кандидат технічних наук, Національний технічний університет «Харківський політехнічний інститут», старший викладач кафедри інтелектуальних комп'ютерних систем; тел.: (093) 083 261 3; e-mail: svetapetrasova@gmail.com.

Петрасова Светлана Валентиновна – кандидат технических наук, Национальный технический университет «Харьковский политехнический институт», старший преподаватель кафедры интеллектуальных компьютерных систем; тел.: (093) 083 261 3; e-mail: svetapetrasova@gmail.com.

Petrasova Svitlana Valentynivna – Candidate of Engineering Sciences (Ph. D.), National Technical University "Kharkiv Polytechnic Institute", Senior Lecturer at the Department of Intelligent Computer Systems; tel.: (093) 083 261 3; e-mail: svetapetrasova@gmail.com.

Кузьміна Марія Олександрівна – Національний технічний університет «Харківський політехнічний інститут», студентка кафедри інтелектуальних комп'ютерних систем; тел.: (073) 078 127 7; e-mail: m.larina@yahoo.com.

Кузьмина Мария Александровна – Национальный технический университет «Харьковский политехнический институт», студентка кафедры интеллектуальных компьютерных систем; тел.: (073) 078 127 7; e-mail: m.larina@yahoo.com.

Kuzmina Maria Oleksandrivna – National Technical University "Kharkiv Polytechnic Institute", Student at the Department of Intelligent Computer Systems; tel.: (073) 078 127 7; e-mail: m.larina@yahoo.com.

Мануйлов Ілля Олександрович – Національний технічний університет «Харківський політехнічний інститут», студент кафедри інтелектуальних комп'ютерних систем; тел.: (099) 451 417 2; e-mail: banger@ukr.net.

Мануйлов Илья Александрович – Национальный технический университет «Харьковский политехнический институт», студент кафедры интеллектуальных компьютерных систем; тел.: (099) 451 417 2; e-mail: banger@ukr.net.

Manuilov Ilya Oleksandrovych – National Technical University "Kharkiv Polytechnic Institute", Student at the Department of Intelligent Computer Systems; tel.: (099) 451 417 2; e-mail: banger@ukr.net.