Юрій Чемерис
Національний університет «Львівська політехніка»

ПРОБЛЕМА АВТОМАТИЧНОГО ОПРАЦЮВАННЯ НЕСТАНДАРТНОГО ПРАВОПИСУ В УКРАЇНСЬКОМУ КОРПУСІ: НАПИСАННЯ РАЗОМ, ОКРЕМО І ЧЕРЕЗ ДЕФІС У ТЕКСТАХ ЖЕЛЕХІВКОЮ

У статті описано проблеми, пов'язані з обробленням україномовних текстів, написаних західноукраїнським варіантом правопису (желехівкою) у період з 1880 до 1910 р. Желехівка мала значну варіативність, що показано на прикладі варіантів написань найчастотніших слів, частково спричинених впливом польського правопису. Це значно ускладнює завдання автоматичного опрацювання старих західноукраїнських текстів.

***Ключові слова**: український правопис, желехівка, ГРАК, варіантність, нормалізація.*

The article describes the problems associated with the processing of Ukrainian-language texts written in the Western Ukrainian spelling (zhelehivka) in the period from 1880 to 1910. Zhelekhivka was highly variable, as shown by the spelling variations of the most frequent words, partly under the influence of Polish spelling. This greatly complicates the task of automatic processing of old Western Ukrainian texts.

***Key words**: Ukrainian spelling, zelekhivka, GRAK, variation, normalization.*

У межах лінгвістики існує й надалі проводиться багато досліджень україномовних текстів різного часу. У 1880–1910 роках існували правописні варіанти, що відрізняються від сучасного українського правопису, зокрема правопис Желехівського, що

був поширений у Західній Україні. Корпусне опрацювання історичних текстів, написаних несучасним правописом, потребує етапу нормалізації, яку здійснюють, наприклад, в історичних корпусах польської мови [5, pp. 3854–3856], після чого до старих текстів можна застосувати інструменти для аналізу стандартного тексту. Желехівка у ГРАКу [6] нормалізована тільки частково, на основі чотирьох правил, розмітка корпусу не враховує всіх особливостей правопису. Для покращення опрацювання текстів желехівкою в ГРАКу потрібен детальний опис цього правопису і наявних у ньому варіантів. Отже, насамперед доцільно виокремити поширені варіанти написання слів та сформувати правила нормалізації для регулярних правописних відповідників.

У межах цього дослідження ми проаналізували нестандартні варіанти написання найчастотніших слів — службових частин мови, займенників, дієслова *бути* — у корпусі текстів желехівкою в ГРАКу-16. Для цього було сформовано підкорпус текстів желехівкою (обсяг 6,5 млн токенів) і виведено частотний список нерозпізнаних слів без урахування власних імен та назв (за CQL-запитом [tag="unknown"&word="[а-яіїєґ].*"]). Обсяг нерозпізнаних слів склав 52 160 одиниць. Далі було відібрано найчастотніші нерозпізнані слова, правописні варіанти яких об’єднано в групи. Нерозпізнані слова з праць Михайла Грушевського не враховувались, оскільки ці тексти містять численні цитати з історичних джерел, що ускладнює пошук.

Сформовано такі групи нерозпізнаних слів, які є правописними варіантами, з різною частотою слів та розміром груп: ***з(із), єсли, є, что, що, ще, між, будь, ли, по, ось, то(от, те, та), ж(же), хоч, ин, бо, би(б), тільки, де, але, для***. Частково ці варіанти могли утворитися під впливом польського правопису, тож польські слова додаємо для наочного порівняння.

1. **з(із)** — *из (345), ізміж (31), изо (31), зперед (20), зпосеред (13)* [śpośród], *зпроміж (9)* [spomiędzy], *іс (9), зь (9), со (7),*

ізпід (5) [spod], *зпоза (4)* [spoza], *из-за (4)*, *з-по-за (2)*, *зачерез (2)*, *ізпоміж (1)*, *де-з (1).*

2. **єсли** — *єсли (312)* [jeśli], *сли (52)*, *єсли-б (15)*, *єсли-би (9)*, *если (7)*, *еси (5)*, *ежели (2)* [jeżeli].
3. **є** — *єсьмо (21)*, *єсмь (8)*, *єсь (7)*, *єсте (5)*, *єсьм (4)*, *єст (2).*
4. **что** — *што (18)*, *что (13)*, *чтоб (2)*, *чтош (1)*, *чож (1).*
5. **що** — *де-що (60)*, *що-ино (9)*, *мало-що (5)*, *що-раз-то (3)*, *за-що (2)*, *що-в (1)*, *що-аж (1)*, *що-йно (1)*, *так-що (1)*, *не-знати-що (1).*
6. **ще** — *єще (56)* [jeszcze], *еще (8)*, *аще (5).*
7. **між** — *міжь (22)*, *мїжь (18)*, *между (3)*, *міжо (3)*, *міжтим (2)*, *по-міжь (1).*
8. **будь** — *небудь (219)*, *будьто-би (52)*, *будьто (44)*, *будьтоби (5)*, *будьтеж (1)*, *будь-щобудь (1)*, *будь-то-би (1)*, *де-будь (1).*
9. **ли** — *ли (23)*, *или (12)*, *што-лї (1).*
10. **по** — *по-за (67)*, *по-при (56)*, *под (6)*, *спопід (2)* [spopod], *спонад (2)* [sponad], *понадь (1).*
11. **ось** — *ось-що (9)*, *ось-як (6)*, *ось-які (2)*, *ось-якої (2)*, *ось-який (2)*, *ось-яке (2)*, *ось-такі (2)*, *ось-така (2)*, *ось-то (1)*, *ось-якою (1)*,*ось-такій (1)*, *ось-таку (1)*, *ось-такого (1)*, *ось-такиб (1)*, *ось- таке (1).*
12. **то(от, те, та)** — *тоє (30)*, *про-те (22)*, *тот (13)*, *про-то (3)*, *то-та (2)*, *томуто (1)*, *то-то-то (1)*, *от-іще (1)*, *от-що (1)*, *не-то (1)*, *такото (1)*, *от-які (1)*, *от-тутечки (1)*, *от-тут (1).*
13. **ж** — *однакож (164)* [jednakże], *ажь (54)*, *все ж (47)*, *ож (30)*, *вжеж (24)*, *тогож (20)* [tegoż], *оже (12)*, *хтож (11)* [któż], *чомуж (11)* [czemuż], *деж (11)* [gdzież], *тоїж (8)*, *жь (6)*, *туж (4)*, *якіж (4)* [jakiż], *яж (4)*, *уж (3)*, *иже (3)*, *такіж (2)* [takiż], *тійже (2)* [tejże], *йогож (2)* [jegoż], *хібаж (2)*, *тіж (2)*, *то-жь (2)*, *такимиж (2)*, *тимже (1)*, *тутже (1)*, *такогож (1)*, *нехайже (1)* [niechże], *такімже (1)*, *опісляж (1)*, *тойже (1)*, *якож (1)*, *такояж (1)*, *длячогож (1)*, *атакож (1)*, *аж'тоді (1).*

14. **хоч** — *хочь (43)* [choć], *хотьби (14), хочь-би (8), хочаби (2), хочьби (1)* [choćby].

15. ин — *ино (20), инак (3).*

16. **бо** — *албо (1)* [albo], *убо (1).*

17. **аби** — *не-аби (2), абим (2), аби-сьте (1), аби-сте (1), аби-в (1).*

18. **тільки** — *нетільки (66), только (10).*

19. **де** — *денекотрі (1), денекотрих (1), денеде (1)* [gdzieniegdzie], *деяку (1), де-якої (1), де-якими (1), де-яка (1), де-чому (1), де-неде (1), де-колидесь (1), де-в-чім (1), де-в-чому (1), де-будь (1).*

20. **але** — *а-ле (3), алекий (1).*

21. **для** — *длятого (341)* [dlatego], *длячого (42)* [dlaczego], *длятого (3), адля (1).*

Також у корпусі наявні нерозпізнані слова, що не ввійшли до жодної групи: *лишь (147), кромі (96), як-раз (69), наконець (69), ід (51), прецї (26), кріз (25), вь (13), себ (10), просторонь (9), оскільки (3), такта (1), о-поки (1), бонема (1), а-тепер (1).*

Подібно до польської мови частка **би(б)** може писатися разом з попереднім словом: *чинитьби (31), булиб (12)* — byłyby, *наколиб (8), булоби (5)* — byłoby, *булаби (5)* — byłaby, *хтоб (1), якіб (1).*

У текстах із желехівкою синтетичні форми майбутнього часу пишуться окремо: *знати му, ходити меш.* Хоча є також випадки написання разом: *-меш (413), -меться (37), -муться (27), -мешся (27), -сьмо (27), -мусь (25), -сьте (23), -метесь (22), -сь (10), -ються (1), -ють (1).* Трапляються випадки, коли така частка стоїть у препозиції («Доки ти — каже, *меш* мене надсажувати роботов тай доки *меш* мене так гидно годувати?»).

Для ефективнішої роботи доцільно нормалізувати подібні випадки, що буде наступним кроком до автоматизації роботи у подальших лінгвістичних дослідженнях. Ще однією актуальною потребою є створення нових інструментів оброблення текстів у корпусі, що прискорить і полегшить роботу дослідника.

Список використаних джерел

1. Барчук М. Фонетико-правописні особливості вокалізму в літературній мові Галичини середини XIX ст. *Мовознавчий вісник*. 2012. Вип. 14–15. С. 27–36. URL: http://nbuv.gov.ua/UJRN/Mv_2012_14-15_6 (дата звернення: 16.05.2023).
2. Білявська Т. Проблема введення нової правописної системи в Західній Україні на сторінках часопису «Зоря» (1880–1897 рр.). *Науковий вісник Херсонського державного університету. Сер.: Лінгвістика*. 2014. Вип. 21. С. 10–15. URL: http://nbuv.gov.ua/UJRN/Nvkhdu_2014_21_4 (дата звернення: 16.05.2023).
3. Gruszczyński W. et al. The Electronic Corpus of 17th- and 18th-Century Polish Texts. *Language Resources and Evaluation*. 2022. Vol. 56. Pp. 309–332. https://doi.org/10.1007/s10579-021-09549-1
4. Kieraś W., Komosińska D., Modrzejewski E., Woliński M. Morphosyntactic Annotation of Historical Texts. The Making of the Baroque Corpus of Polish. Text, Speech, and Dialogue. *Lecture Notes in Computer Science*. 2017. Vol. 10415. URL: https://link.springer.com/chapter/10.1007/978-3-319-64206-2_35
5. Kieraś W., Woliński M. Manually annotated corpus of Polish texts published between 1830 and 1918. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. 2018. URL: https://www.researchgate.net/publication/325217041_Manually_annotated_corpus_of_Polish_texts_published_between_1830_and_1918 (date of access: 16.05.2023).
6. Генеральний регіонально анотований корпус української мови (ГРАК): вебсайт. URL: uacorpus.org (дата звернення: 16.05.2023).