Normalization of Early Modern Ukrainian in GRAC: the Case of Lesia Ukrainka's Works[1]

Maria Shvedova[1], Nataliia Prydvorova[2] and Ilona Skibina[3]

[1, 2,3] *Lviv Polytechnic National University, Bandera Str., 12, Lviv, 79000, Ukraine*

Abstract

The paper deals with the representation of the oeuvre by Lesia Ukrainka (1871-1913) in the General Regionally Annotated Corpus of Ukrainian (GRAC). The poet's texts offer numerous challenges with regard to orthographical and linguistic variation, and existing editions feature distinct strategies with regard to their normalization, often contradictory and incoherent. The authors propose different rule-based patterns that enable more efficient processing of such texts within the corpus. The approach is relevant not only for Lesia's works but for a wider range of Ukrainian texts of the period characterized by similar features.

Keywords

Ukrainian language, orthography, normalization, rule-based annotation, idiolect

1. Introduction

GRAC is a reference corpus of the modern Ukrainian language [1]. According to tradition, the history of the modern Ukrainian language starts from the poet and dramatist Ivan Kotlyarevsky, that is, from the late 18th - early 19th centuries. However the majority of GRAC is composed of contemporary Ukrainian texts. One of the reasons for this disbalance is the imperfection of available tools for processing Ukrainian texts in their historical spellings. Ukrainian orthography until 1928 had no unified standard. In particular many variants of spelling existed in the 19th century. It is because of spelling differences that the system of morphological analysis used in GRAC leaves many unrecognized words and grammatical forms while processing such texts. This system is based on the VESUM electronic dictionary of Ukrainian [2].

For this reason, much of the 19th- and early 20th-century texts have been added to GRAC in a modernized orthography according to later publications rather than in their original spelling. But using such editions in the corpus is also problematic because publication of historical texts, particularly those of the Soviet period, were not only normalized in terms of spelling and grammar, but often considerably edited and censored [3, 4]. Hence these editions are not a comprehensive source of information on the linguistic period when the text in question appeared. In addition, there are many texts published in the 19th and early 20th centuries and never reprinted thereafter (e.g., newspaper texts, the bulk of magazine publications, book editions). Therefore, there is a need to connect additional tools to the existing system of morphological analysis in order to process the old spellings. So far, VESUM has been supplemented with rules for the recognition of old and non-standard spellings common in the corpus, regardless of spelling. Besides, a separate module was created for the recognition of zhelekhivka (the spelling used in Western Ukraine in the 1880s-1920s), where several basic rules for processing texts written in this system were additionally applied [4].

Adding non-standard spellings and grammatical forms to the common dictionary is not always a good solution, because it can increase grammatical homonymy. For example - the token ма in older texts is usually a verb: *Хто дба, той ма (Номис, Українські приказки, прислів'я і таке інше, 1864),*

[1] The work was partially supported by the visiting program of Polish Academy of Sciences at the Institute of Computer Science PAS

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland

EMAIL: mariia.o.shvedova@lpnu.ua (M. Shvedova); nataliia.prydvorova.mflpl.2021@lpnu.ua (N. Prydvorova); ilona.skibina.mflpl.2021@lpnu.ua (I. Skibina)

ORCID: 0000-0002-0759-1689 (M. Shvedova); 0000-0002-8305-3479 (N. Prydvorova); 0000-0003-0291-0414 (I. Skibina)

whereas in modern texts it is most often a colloquial form of address to one's mother (a vocative): — *Ма, вибач, я не знав… (Ірен Роздобудько, Арсен, 2012).* To resolve such ambiguity one needs separate modules within the grammar dictionary for separate groups of texts.

This paper presents the first experience of creating an additional VESUM module for processing an author's corpus of texts. This module preserves multiple cases of old and individual spelling. It consists of normalization algorithms used in lemmatization of non-standard tokens that do not affect the appearance of the text in the corpus. Non-standard spelling and grammatical variants not covered by these algorithms were added to the general dictionary, provided that they were frequent in the corpus texts. Thus the accuracy of morphological annotation in the texts by Lesia Ukrainka (1871-1913) has been slightly improved.

The first part of the paper deals with the characteristic features of the spelling of Lesia Ukrainka and with textual principles in different editions of her oeuvre. The second part includes analysis of how her spelling is rendered in the latest complete collection of works [5]. The third part of the article is dedicated to the issues behind the morphological analysis of these texts. Rules are proposed to improve the recognition of non-standard words.

2. Characteristic of the Lesia Ukrainka's spelling

The main critical editions of Lesia Ukrainka's works give different versions of the texts. The most standardized one, and also the closest to modern spelling norms, is the academic edition of the 1970s in 12 volumes [6], which used to be a standard edition for a long time.

Many more specific features of Lesia Ukrainka's language are preserved in modern editions [7-9, 5], although the principles of textual normalization in these series are somewhat different. In particular, the 2016-2018 edition of Letters unified the spelling of borrowings, applying the "rule of nine" present in the contemporary norm: *сюрпріз - сюрприз, цітата - цитата, діктувати - диктувати,* etc. [10]. In CW-14 the author's (variant) spelling of such words is preserved. On the other hand, the editors of [7-9] preserve the orthographic variants that do not affect the pronunciation (e.g., *пьять, моеї, такоі, тут таки, було-б, де-що, щож*), while in CW-14 such cases are rendered in the modern spelling.

Here is an example of a comparison between the text of Lesia Ukrainka's letter of 6 July 1889 to her mother in CW-12 and CW-14. We can see that CW-12 uses relatively stronger standardized spelling (*такій - такий*), grammatical forms (*світа - світу*), lexical variants (*куповать - купувать, нанять - найнять, завдатку - задатку, тілько - тільки*). In the CW-12 edition multiple passages were omitted due to censorship. The editors of [7-9] restore these and mark them in italics [7].

Lesia Ukrainka's own spelling was not well-established, as noted, for many lexemes different spelling variants are attested. In addition, she used different spelling systems. For example, as noted in the commentary on CW-14, a draft of the poem "Одержима"/"Obsessed" was written in kuleshivka, and while rewriting of the original text, zhelekhivka was used [5]. In her correspondence Lesia Ukrainka sometimes used dragomanivka [7].

In addition, Lesia Ukrainka added individual features to her spelling. The correspondence of Lesia Ukrainka with Ivan Franko in 1892, before the publication of her poetic collection "На крилах пісень"/"On the wings of songs", shows that the attitude to spelling was quite free at that time (not unlike the attitude to the author's punctuation nowadays). Lesia Ukrainka writes: "The word *чі* let it be, because we got used to it, that's our pronunciation", "I do not stand for the letter *ї*, but it is right just like other iotated letters: *ю, я, є* etc; I stand for true phonetic spelling (a radical one), but as I cannot use this one in print, then I should deal with that one somehow. After all, I am also opposed to that *ї* at some way, you can replace it with *йі*, or just - *і*, such as *своі, Україна*, etc.", "The word *слізьми* really should be written with a soft sign, because it is pronounced that way. The word *вьяне* can be written either so, or *вйане*, but in no way *вяне*, for that, in fact, will get the Russian pronunciation. *Инший* let remain so too" [5]. From this we see that the orthographic standard was not so strict, variations were allowed and discussed, and that it was very important for Lesia Ukrainka to reproduce accurately the everyday language. The language of Lesia Ukrainka reflected some orthoepic, grammatical, and lexical features of the Western Polissia dialect [10], and this not only in her works of fiction, where she used them for stylistic purposes, but, for example, in her letters as well.

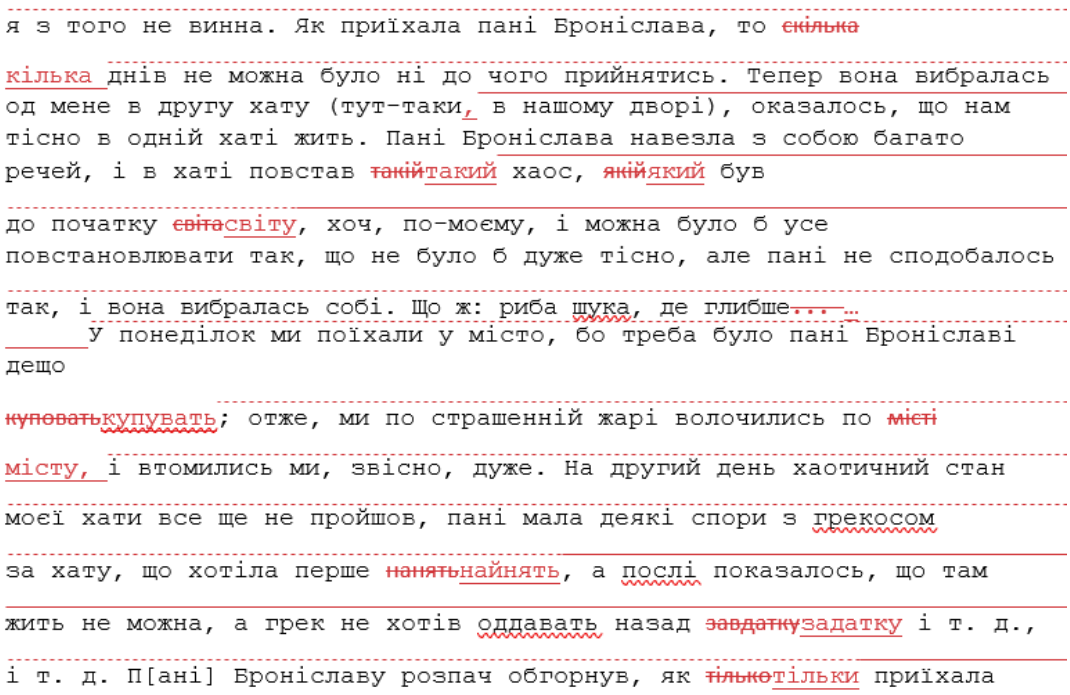

Figure 1: Comparison of the text of Lesia Ukrainka's letter to her mother (July 6, 1889, Odesa) in CW-12 (underlined) and CW-14 (crossed out)

In the texts of CW-14 in many cases the original spelling variation of Lesia Ukrainka is preserved that reflected the pronunciation variants. In the texts of CW-14 the parallel use of orthoepic variants *чи/чі, квартира/квартіра, стаття/стать, завжди/завжді, теперешній/теперішній, матерьял/матеріал, настілько/настільки, отель/готель, подлий/підлий, поворіт/поворот, почта/пошта*, etc. is attested. In the texts of CW-12, as indicated in the notes to the publication of the letters, such cases, where the spelling of the author was not uniform, are rendered according to the modern norms [6].

There are also some pairs of lexical variants (absolute synonyms) in Lesia Ukrainka's texts. Some of these foci of variation did not survive into the contemporary literary language: *бабушка/бабуся, іюнь/червень* (and also other names of months), *одкритка/листівка, случай/випадок, поводіння/поведінка, устроїтися/влаштуватися*, etc. Only the second member of each pair is attested in the present-day standard.

Such features of Lesia Ukrainka's language are important not only for the study of her individual style but also as a sample of the Ukrainian language of this period. Thus, such variants as *тільки, скільки*, etc. that were used by Lesia Ukrainka but are now non-normative, were used in the language of her contemporaries, members of her family [11]. The corpus of Lesia Ukrainka's texts marks a certain stage of formation of the norm. It is therefore valuable for the historical part of a general linguistic corpus.

3. Text normalization in the 2021 edition

The 2021 edition (CW-14) was chosen for the GRAC corpus, as it reproduces the language of Lesia Ukrainka and her time more accurately, a feature that is important for a reference-type corpus. The texts of Lesia Ukrainka according to the CW-12 version, which had been contained in the corpus until version 13, in GRAC.v.14 (compiled 4.11.2021) were replaced by the texts from CW-14.

It is known that Lesia Ukrainka herself paid much attention to the written form of her phonetics and morphology, such as the palatalized pronunciation [ч]/[ch] in the word *чі*, non-palatalized adjective endings in the nominative plural (*весняні, молодії*), non-palatalized pronunciation of final hushing sibilants (*між, хоч, ще ж*), etc. [12].

All texts published in CW-14 were added to GRAC.v.14. The Lesia Ukrainka corpus consists of 1,225 texts and has a total size of almost 1.5 million tokens, of which about 1.3 million tokens of original works and 0.2 million tokens of translations; 0.56 million tokens of letters, 0.33 million tokens of poetry, 0.23 million tokens of fiction, 0.15 million tokens of journalism, 0.12 million tokens of historical studies, and 0.08 million tokens of folklore records (plays in verse and prose are counted as part of poetry and fiction respectively).

According to the editors' explanations in the preface to CW-14, the texts are rendered linguistically according to the following basic principle. In many cases, it is necessary and possible to adapt the texts to the modern spelling (e. g. yaryzhka and drahomanivka spelling systems are transliterated to the present-day conventions), but this procedure should not change the author's phonology, morphology, and lexicon. The lexicon was preserved as well as inflectional, derivational and phonetic variants that were habitual for the literary standard of that time and/or for the language of Lesia Ukrainka, irrespective of their status in the contemporary Ukrainian language. If the writer herself allowed different spellings of words even within a single text, then a uniform variant could be chosen – either the one that has a greater frequency in the text or the one attested later within the same work; such cases and the editor's choice of a variant are specially commented [13].

So, the texts of Lesia Ukrainka in the edition of 2021 are partially normalized. But the principles of normalization chosen by the editors are not always clear. In some cases the editors are to choose a variant at the level of a single text (either the more frequent one or the one attested later), that is, at the level of the whole edition, the choice of such variants was not uniform and may have been made differently in different works.

We see that in the 2021 edition as a whole, the proclaimed normalization principles were not always implemented consistently. For example, according to the principles of the edition, the words spelt together, separately, and hyphenated, while this does not affect pronunciation, should be normalized. However spellings that are not conform to the modern standard are attested in the edition: *Від змішаня двох різних критеріїв, релятивного й абсолютного, д. Ганкевич не може собі дати ради з псіхолопією новітніх християн і нехристиян і* ***де-далі*** *все більше перемішує не тілько псіхолопічні, а навіть хронолопічні моменти. Так, з поводу того, що українці плачуть тепер (1903 р.) над недолею Бурів, д. Ганкевич питає, де вони були, коли царат давив геройську польську революцію,* ***себ то*** *1863 р.? (Френсіс Артур Фегі ● 1906 ● Справа ірландськоі мови ● пер. Леся Українка). Коли російські революціонери часом користають з імення, пророчих візий та* ***де-яких*** *афорізмів Толстого, то се така сама «іронія історії», як і те, що політик соціал-демократ бере собі motto у безполітичного анархіста часів першого християнства, та щей називає того «непротивленця» якобінцем! (Леся Українка ● 1903 ● Замітки з приводу статі «Політика і етика»).*

The spelling of the soft sign and apostrophe is not always normalized either: *Та людина, що візьме на себе се завданє, почне з того, що відкине при перегляді все* ***сьвітське****, театральні твори, не-релітійну поезию, політичну історию і т. і. (Моріс Верн ● 1894 ● Біблія або книги Старого Завіта ● пер. Леся Українка). Здумайте, до чого мало було часу — я тілько вчора вперше могла заграти як годиться! хоч* ***фортепьяно*** *вже з тиждень як привезене (Леся Українка ● 1899 ● Лист до О. Ю. Кобилянської 16-17 жовтня 1899 р. Київ).*

The treatment of the voiceless and voiced consonants, which, according to the editorial instructions, should be reproduced according to the modern norm, is not always consistently normalized: *На душі стало* ***лехко****, так, як бувало після довгої молитви в нашій сільській церковці (Мендель Розенбаум ● 1902 ● Великдень у турмі ● Леся Українка). Послухайте, сеньйоре де Маранья, я вас не встигла* ***роспитати*** *вчора (Леся Українка ● 1911 ● Камінний Господарь). Нехай ця безкрая надія непевна, але ж хиба* ***роспач*** *певніший? (Леся Українка ● 1906 ● Утопія в белетристиці).*

The inconsistency revealed in Lesia Ukrainka's own spelling, as well as in the modern edition of the texts (2021) must be taken into account in the morphological analysis of the texts in the corpus.

4. Rules proposed to improve the recognition of non-standard words

A system of morphological analysis based on the VESUM grammatical dictionary is used for morphological markup of the GRAC corpus [2]. This system is designed for the modern standard,

although it uses separate tools to lemmatize texts written using other orthographic systems (zhelekhivka, skrypnykivka, the Soviet spelling of 1933, the spelling of 1992, the modern spelling of diaspora, etc. [18]) and some of their elements. Some of the frequent non-standard spellings are added to the dictionary, whereas others are recognized using dynamic tagging tools [4].

While annotating the texts of the 19th - early 20th century, GRAC does not apply transliteration rules to all the texts before morphological markup, as this is the case for the historical corpus of Polish [14, 15] or for the historical corpus of Russian [16, 17]. The morphological analysis program annotates the texts in their original spelling and applies the rules of dynamic tagging to the unrecognized words.

The available tools (a dictionary containing the most common nonstandard variants, plus the rules of dynamic tagging) allow us to recognize some cases of nonstandard spelling preserved in the texts of the CW-14, such as the spelling of the letter ґ instead of the modern г (*єґер, енергія, оргія, релігійний, телеґрама, фонограф, леґенда, трагедія, етнографічний,* etc.), some words with the initial и- (*инший, инакше*), alternative forms of the nominal genitive singular with the ending -и (*пам'яти, імени*), phonetic treatment of borrowings that departs from the most frequent modern variant: with soft [l'] (*сальон, кляса, плян, лямпа, клясовий, мельодія, пляц, Філярет, Голяндія, фільософічний, скарлятина, реклямувати, платформа, парлямент, новеля, лявіна, кільо, кольоніст, демонольогія, деклямувати, баляда*), words with -ія-, -'я- instead of normative -іа- (*матеріял, матеріяльний, азіятський, азіят, варіянт, матер'ял, фіялка, спеціяльний, соціялізм, соціялістичний, соціяльний, індіянка, уніятський*), with the final -тер instead of the normative -тр (*міністер*), derivational variants (*роля, заля*), as well as some other orthoepic, word-formation, grammatical and orthographic variants (*ріжний, ріжниця, ратунок, житє, соняшний, претенсія, европейський, Европа, европеєць , струмент, мущина, сальоновий, бенькет, багацтво, альфабет; люде; нераз, неначеб, аби-який, їйбогу,* etc.). Such cases are either added to the dictionary as alternatives (with the :alt tag) or fall under one of the rules of dynamic tagging (Andriy Rysin) described in [4].

The system also recognizes and correctly lemmatizes abbreviations with full forms presented in square brackets which is traditional for critical editions. *Се я пробую, чи не віддасть він мені тиї 200* ***р[ублів]****, що 5 літ тому взяв (Леся Українка ● 1911 ● Лист до О. П. Косач (сестри) 14 грудня 1911 р. Хоні. CW-14).*

But, in addition, the texts of CW-14 retain some specific features of the language and spelling of Lesia Ukrainka, which are not recognized by the program of morphological analysis based on VESUM. There are 55 593 unrecognized tokens, or 4.89% of the total sized of the subcorpus, which far exceeds the number of words usually left unrecognized in the analysis of standard texts.

Table 1

The result of the morphological analysis system based on VESUM

Texts of Lesia Ukrainka from the 2021 edition	Modern standard texts
Known: 1084367, unknown: 55593, 4.9% Known unique: 95709, unknown unique: 23460, 19.7%	Known: 646616478, unknown: 12074958, 1.8% Known unique: 2791389, unknown unique: 3466095, 55.4%

An additional list of dynamic tagging rules was created to lemmatize some of these words, which are used only for the analysis of the Lesia Ukrainka corpus.

The rules were formulated based on the analysis of the corpus of Lesia Ukrainka's texts from the 2021 edition (CW-14). They cover: orthographic and word-formation variants, grammatical variants (endings), common irregular variants for some word forms.

4.1. Orthographic and word-formation variants

1. .*йі.* => .*ї.*
   Although the editors have generally normalized the old orthographic variants without affecting pronunciation, some texts retain spellings with -йі- (*йіх, йім, йій, свойій, звичайів, Євреїі,* etc.)
   Similarly:
   .*іі.* => .*ії.*
   (*Сиріі, націі, історіі, Данііл, Єзекііл, Манассіін,* etc.)
   .*оі.* => .*ої.*
   (*своіх, божоі, цілоі, римськоі, малоі, давньоі, історичноі,* etc.)
2. .*і.* => .*и.*
   In many words the author's spelling is preserved with the *і* letter in place of the standard *и* (*чі, завжді, назавжді, звідсі, почасті, чотирі, дінастія, стіль, асірійський, Тіфліс, цівілізація, історік, крітік, сімпатія, квартіра, тіф, збірання, режім, крівавий, вмірати, блакітний, кредітор, трівога, мотів, умірати,* etc.).
3. .*иї => .*ії
   Endings of adjectives in the nominative plural (e.g, *тиї, любиї, чорниї, білиї, ясниї, золотиї, молодиї, весняниї, німиї, темниї, срібниї, святиї, палкиї, малиї, крівавиї, зеленіїї, дорогиї, добриї, високиї, буйниї, широкиї, стариї, новиї, живиї, ворожиї, чудовиї, тихиї, таємниї, страшниї, смутниї, рясниї, нічниї, непевниї, мудриї, людськиї, жовтиї*) and of some nouns (e.g, *нації, партиї, організації, Франціїї, цівілізації, фікції, фантазиї, рації, пунктуації*).
   The VESUM-based morphological analysis system already recognizes non-standard long forms of adjectives with -і-endings (for example: дрібнії adj:p:v_naz:compb:long) [2], so the suggested replacement will be sufficient for lemmatizing adjectives of this type (*зеленіїї, дорогиї, добриї*).
4. .*рь => .*р
   The soft sign after р on the end of masculine nouns in the initial form (*царь, лікарь, лицарь, владарь, господарь, олтарь, крамарь, вихорь, пузирь, кобзарь, звірь, шинкарь, Цезарь, писарь, ліхтарь,* etc.) and in other words after the final р: *бурь, вірь, матірь, теперь*.
5. .*ійш.* => .*іш.*
   Adjectives and deadjectival adverbs with the suffix -ійш (*пізнійше, ранійше, скорійше, найчастійше, міцнійше, труднійше, простійше, певнійше, найяснійший, докладнійше, ощаднійше, найстрашнійший, найсвятійший, найпотрібнійший, найвірнійший, цікавійший, пильнійший, новійший, найславнійший, найповнійший, найміцнійший, найвидатнійший,* etc.).
6. .*стн.* => .*сн.*
   Words with the -стн- combination: (*перстні*, *первістний*, *користно*, *намістник*, *безучастно*, *устний*, *розпустний*, *пристрастно*, *зловістний*, *честний*, *хрестний*, *провістник*, *непричастний*, *ненавистний*, *напастник*, etc.).
7. и.* => і.*
   Words with initial и- (*именно, искра, иньший, имення, инак, император, инде, идолянин, испанський, играшки,* etc.)
8. .*кілько=>.*кільки
   .*тілько=>.*тільки
   Words кілько, тілько and their derivates (*тілько, скілько, стілько, наскілько, настілько, оскілько, остілько, ніскілько, хтозна-скілько, скілько-небудь*)
   Such variants with the ending -о were characteristic for Lesia Ukrainka; according to the research of S. Bohdan variants in -о quantitatively prevail in the corpus of her letters: *стілько – стільки* 131:6, *настілько – настільки* 53:5, *скілько – скільки* 178:6, *наскілько – наскільки* 59:2 [11].

4.2. Variant forms

1. .*не => .*ни
   (*єгиптяне, християне, галичане, римляне, вавілоняне, самаряне, росіяне, магометане, англічане, кияне, асіро-вавілоняне, ізраельтяне, правдяне, островитяне, мусульмане)*
2. .*ре => .*ри
   (*бояре, болгаре*)
3. .*ів => .*ей
   (*прикростів, деталів, постатів, розповідів, тінів, паралелів, подорожів, неприємностів, національностів, слабостів, повинностів, заповідів, відповідів, умілостів, капітелів, знаменитостів, галузів, банальностів, індівідуальностів, єресів*)
4. .*ннів => .*нь
   (*створіннів, повстаннів, питаннів, зібраннів, вприскуваннів, виданнів, почуваннів, оповіданнів, вітаннів, бажаннів, ученнів, порівняннів, порученнів, пориваннів, переконаннів, обливаннів, непорозуміннів, нагадуваннів, змаганнів, вимаганнів*)
5. .*ддів => .*дь
   (*знаряддів, привиддів*)
6. .*ттів => .*ть
   (*понять, століть*)
7. .*ччів => .*ч
   (*обличчів*)
8. .*ов => .*ів
   (*голов, гріхов*)
9. .*ови => .*ові
   (*фабрикантови, робітникови, народови, впливови, батькови, урядови, духови, братови, богови, Соломонови, Авраамови, чоловікови, хистови, флейтистови,* etc.)
10. .*жу => .*джу
    (*хожу, сижу, ненавижу, знахожу, вихожу, сужу, проважу, поражу, зражу, догожу, спроважу, лагожу, углежу,* etc.)
11. Word forms: мні => мені, него => нього, ви-те => ви, сею => сією, сего => сього, теї => тієї, меї => моєї, свеї => своєї, єі => її, племени => племені, людий => людей, близше => ближче, легче => легше

Applying these basic rules helped to improve the result of morphological markup:

Table 2

The result of the morphological analysis of Lesia Ukrainka texts from the 2021 edition

Without applying additional rules	After applying additional rules
Known: 1084367, unknown: 55593, 4.9% Known unique: 95709, unknown unique: 23460, 19.7%	Known: 1091052, unknown: 48908, 4.3% Known unique: 97237, unknown unique: 21932, 18.4%

Adding new rules can help improve these results.

4.3. Variant paradigms

In addition, when analyzing the texts of Lesia Ukrainka, whole variant paradigms with missing alternation in the root were found. They were added as variant paradigms to the main dictionary, hence they are not specific rules only for Lesia's corpus. These are such word forms:

[word="каміню|каміневі|камінем|каміні"]
[word="річі|річей|річах|річами"]
[word="зовуть|зовіть|зову|зови|зовеш"]
[word="зоветься|зовуся|зовуться|зовусь"]

To a number of masculine nouns in the dictionary a non-standard variant of ending -a in the genitive singular was added, which is recorded in the corpus of Lesia Ukrainka texts:

авторітета, виїзда, візіта, всесвіта, гонорара, діагноза, діалога, дуета, журнала, закона, ідеала, клімата, конкурса, культа, курса, легіона, луга, манускрипта, момента, мотіва, народа, овса, пансіона, потока, похода, престола, прецедента, приказа, прінціпа, проєкта, процеса, реферата, рода, романа, романса, сезона, сінедріона, скандала, сна, страха, суда, театра, текста, тома, трактата, тріклініума, урока, уступа, факта, фатума, хамсіна, хаоса, характера, храма, часа, шума.

4.4. Morphological variants of lexemes

A separate observation should be made on the nouns with the root -пис. Lesia Ukrainka used them in two versions, masculine (which is now the normative for them) and feminine: *часопис/часопись*. Feminine variants are recorded in the texts: *часопись, рукопись, допись, літопись, правопись, запись, опись, житєпись*. In the indirect cases for these words, the forms of both paradigms, masculine and feminine, are also recorded in the Lesia Ukrainka corpus. For example, the stem часопис- in the corpus of Lesia Ukrainka texts yields the following forms:

часописі	23
часопись	20
часописях	4
часописів	3
часописи	3
часописей	1
часопис	1

These variants of the feminine gender of nouns with the root -пис, recorded in the corpus of texts of Lesia Ukrainka, are added to the dictionary as separate lemmas.

4.5. Other groups of words suggested for being included into the main dictionary

1. Old names of months of Latin origin: *январь, февраль, март, апріль, май, іюнь, іюль, августъ, сентябрь, октябрь, ноябрь, декабрь*
2. Old abbreviations: *і т. и.* (і таке инше), *д.* (добродій), *гл.* (глава), *до Р. Х.* (до Різдва Христового), *до Хр.* (до Христа), *єв.* (євангеліє), *С.-Д.* (соціал-демократи), *Р. У. П.* (Революційна українська партія).
3. All proper names used in the corpus of Lesia Ukrainka at least twice - 964 lexemes (The complete list of personal names including those recorded once exceeds 5,500).

5. Prospects

The corpus is constantly being updated, in particular with old texts written using different spelling systems. The morphological analysis of the corpus uses a system designed for modern spelling, which is also actively updated with non-standard variants used in many older texts. But such updates do not cover the old spellings and their individual variants sufficiently (as in the case of Lesia Ukrainka). The experience of processing the Lesia Ukrainka corpus has shown that for such subcorpora with different spellings it is advisable to include additional rules to capture individual spelling and grammatical variants. For a closed subcorpus that will not be updated in the future, such as the complete works of an author, it is possible to perform morphological analysis even of all the word forms using manual processing. However, we must bear in mind that in an expanding large historical corpus the number of cases requiring the use of additional rules steadily increases which reduces the efficiency of this method.

6. Acknowledgements

Andriy Rysin for technical guidance and support. Yurii Hromyk, Dmytro Sichinava for the professional advice.

Thanks to the master's students of Lviv Polytechnic National University who helped to prepare and annotate the Lesia Ukrainka's texts for the corpus: Kateryna Sukhar, Andriana Hevalo, Anastasiia Karavan, Olena Horobets, Oksana Kurtiak, Yuliia Kucher, Maryna Lupain, Iryna Ortynska, Diana Rokytska.

7. References

[1] M. Shvedova, R. von Waldenfels, S. Yarygin, A. Rysin, V. Starko, T. Nikolajenko et al. GRAC: General Regionally Annotated Corpus of Ukrainian. Electronic resource: Kyiv, Lviv, Jena. 2017–2022. URL: http://uacorpus.org.

[2] V. Starko, A. Rysin. Velykyi elektronnyi slovnyk ukrainskoi movy (VESUM) yak zasib NLP dlia ukrainskoi movy [Large Electronic Dictionary of Ukrainian (VESUM) as an NLP Tool for the Ukrainian Language], Halaktyka Slova, Vydavnytstvo dim Dmytra Buraho [Dmytro Burago Publishing House] (2020) 135–141.

[3] O. Drul, Popravliuvanyi Franko [Corrected Franko], Zbruch, 2015. URL: https://zbruc.eu/node/35977

[4] M. Shvedova, A. Rysin, V. Starko, Handling of Nonstandard Spelling in GRAC. 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), Vol. 2, Lviv, Sept. 22-25 2021, pp. 105-108. doi: 10.1109/CSIT52700.2021.9648834. URL: https://ieeexplore.ieee.org/document/9648834

[5] CW-14: Lesia Ukrainka. Povne akademichne zibrannia tvoriv: u chotyrnadtsiaty tomakh [Full Collection of Works in 14 volumes], Volynskyi natsionalnyi universytet imeni Lesi Ukrainky [Lesya Ukrainka Volyn National University], Lutsk, 2021. URL: https://ubi.org.ua/uk/activity/zibrannya-tvoriv-lesi-ukra-nki-u-14-i-tomah

[6] CW-12: Lesia Ukrainka. Zibrannia tvoriv u 12 t. [Collection of Works in 12 volumes], Naukova dumka, Kyiv, 1975-1979.

[7] Lesia Ukrainka. Lysty: 1876-1897 [Letters: 1876-1897], in: V. A. Prokip (Savchuk) (Ed.), Komora, Kyiv, 2016.

[8] Lesia Ukrainka. Lysty: 1898-1902 [Letters: 1898-1902], in: V. A. Prokip (Savchuk) (Ed.), Komora, Kyiv, 2017.

[9] Lesia Ukrainka. Lysty: 1898-1902 [Letters: 1898-1902], in: V. A. Prokip (Savchuk) (Ed.), Komora, Kyiv, 2018.

[10] I. H. Matviias. Varianty ukrainskoi literaturnoi movy [Variants of the Ukrainian literary language], Instytut ukrainskoi movy NAN Ukrainy [NASU Institute of Ukrainian Language], Kyiiv, 1998, 162 p.

[11] S. Bohdan. Pro «tilko» i ne tilky v movotvorchosti Lesi Ukrainky: u poshukakh idiostyliu [About 'til'ko' and not only in Lesia Ukrainka's linguistic creativity: in search of idiostyle], Kultura slova 93 [The culture of word 93], 2020, pp. 100-114.

[12] L. P. Miroshnychenko. Pro zberezhennia fonetychnoi systemy Lesi Ukrainky u maibutnikh publikatsiiakh yii tvoriv [Preserving Lesya Ukrainka's phonetic system in the upcoming publications of the poetess' works.], volume 8 of Spadshchyna: Literaturne dzhereloznavstvo, tekstolohiia [Heritage: Source Studies in Literature. Textology], Laurus, Kyiiv, 2013, pp. 14–21.

[13] V. Aheieva, Yu. Hromyk, O. Zabuzhko, I. Konstankevych, M. Moklytsia, S. Romanov. (Eds.), Dramatychni tvory (1896–1906) [Dramatic works (1896–1906)], volume 1 of Lesia Ukrainka: Povne akademichne zibrannia tvoriv u chotyrnadtsiaty tomakh [Full Collection of Lesia Ukrainka's Works in 14 volumes], Volynskyi natsionalnyi universytet imeni Lesi Ukrainky [Lesya Ukrainka Volyn National University], Lutsk, 2021, pp. 7-10.

[14] W. Kieraś, D. Komosińska, E. Modrzejewski, M. Woliński. Morphosyntactic Annotation of Historical Texts. The Making of the Baroque Corpus of Polish. In: Ekštein, K., Matoušek, V. (eds)

Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science, vol 10415. Springer, Cham. URL: https://doi.org/10.1007/978-3-319-64206-2_35
[15] W. Gruszczyński, D. Adamiec, R. Bronikowska, et al, The Electronic Corpus of 17th and 18th-century Polish Texts, Lang Resources & Evaluation 56, 2022, pp. 309–332. URL: https://doi.org/10.1007/s10579-021-09549-1
[16] A. E. Poliakov. Problemy i metody analiza russkih tekstov v doreformennoy orfografii [Problems and methods of analysis of Russian texts in pre-reform orthography], series 11 of Kompyuternaya lingvistika i intellektualnyie tehnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii «Dialog» [Computational Linguistics and Intelligent Technologies: based on the materials of the annual International Conference "Dialogue"], Moscow, 2012, pp. 536–547.
[17] A. E. Polyakov, S. O. Savchuk, D. V. Sitchinava, Grammaticheskij slovar' dlja avtomaticheskogo analiza tekstov XVIII–XIX veka: pervye rezul'taty [A grammar dictionary for automatic analysis of the XVIII–XIXth century texts: first results], Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog» (Bekasovo, 29 maja — 2 ijunja 2013 g.) [Computational Linguistics and Intelligent Technologies: Based on the materials of the annual International Conference "Dialogue" (Bekasovo, May 29 - June 2, 2013)], 12 (19), Vol. 1: Osnovnaja programma konferencii [The main conference program], RSHU publishing house, Moscow, 2013, pp. 644-666. URL: https://www.dialog-21.ru/media/1308/dialog_2013_vol1web.pdf
[18] V. V. Nimchuk, N. V. Puriaieva, Istoriia ukrainskoho pravopysu: XVI—XX stolittia [The History of Ukrainian Spelling: 16th to 20th Century], Naukova Dumka, Kyiv, 2004.