

# A Search Engine for the Large Electronic Dictionary of the Ukrainian Language (VESUM)

**Tamila Krashtan**

Lviv Polytechnic National University, Lviv, Ukraine

E-mail: tamila.krashtan@gmail.com

## **Abstract**

This paper presents a new search engine developed for the Large Electronic Dictionary of the Ukrainian Language (also known as VESUM)—a project aiming at generating a morphological dictionary for the Ukrainian language, which is also used in a Ukrainian POS-tagger. The aim of the current project is to set up a more user-friendly interface with broader search options, which at the same time provides more information contained in the Dictionary database. The newly developed search functionality for the Ukrainian Dictionary is built upon the search engine created for the Belarusian grammar database and utilizes grammar tags defined in the VESUM database. It enables the usage of wildcards in search queries and allows a user to set up search grammars. The developed system provides more extensive search options and a way of displaying lemma information that is more structured and transparent both for professionals and non-linguists. It is well-suited for the addition of new tags and search parameters (including, but not limited to, conjugation classes and variations in the orthography of certain words) which will be featured in future versions of the software.

**Keywords:** search engine; online dictionary; Ukrainian; VESUM

## **1. Introduction**

There is a multitude of well-developed NLP tools and databases available for the Ukrainian language that are widely used across various software applications. Nevertheless, the information in such databases is often stored in formats that are not easily usable and are not conveniently consumable by the public.

One such database is the Large Electronic Dictionary of the Ukrainian Language, also known as VESUM, after its Ukrainian acronym (Rysin & Starko, 2005–2023). It is a morphological dictionary that describes the lemmas of the Ukrainian language along with their inflected forms, supplied with grammatical and semantic tags. The data from this dictionary is used in such projects as LanguageTool spellchecker or Wikipedia search (Rysin & Starko, 2020). The creators of the dictionary also provide the data both in the raw format and through a simple search form. However, in order to comprehend the search results, one may need to thoroughly go through pages of documentation, while any search that goes beyond a keyword requires one to create own query scripts to run on the raw data.

The aim of the current project is to build a user-friendly search interface for the VESUM that would leverage most of the data available in the dictionary, including some pieces of information that might be inaccessible through the existing simple search form. This task includes several steps: 1) an analysis of the VESUM structure and the ways to put this structure into queries; 2) an overview of the existing search systems for the dictionaries and grammatical databases of other languages; 3) an implementation of the first versions of the search tool and planning of the future directions of the development. Each of the outlined steps is described in detail in the sections that follow.

## 2. VESUM

VESUM, the Large Electronic Dictionary of the Ukrainian Language, was created in 2005 as a part-of-speech database. Since then, the dictionary itself or its modifications have been utilized in a number of projects, including search engines of Ukrainian Wikipedia and the General Regionally Annotated Corpus of Ukrainian, or GRAC, as well as the Ukrainian spellcheckers in LibreOffice or LanguageTool (Rysin & Starko, 2020).

Since the dictionary is used in spellchecking software, it contains not only those Ukrainian lemmas that are considered a part of literary language, but also corrupted, colloquial, dialectal, and other non-standard forms, each marked accordingly to indicate its usage mode.

The initial version of the dictionary was based on several printed dictionaries of the Ukrainian language (Krytska et al., 2011; Busel, 2005; Karpilovska, 2013), and the database is constantly being updated with new entries, in particular, the untagged words found in the GRAC (Shvedova et al., 2017–2023). The GitHub page of the dictionary (Project to generate POS tag dictionary for the Ukrainian language) contains the latest available version of the database and enables users to make suggestions on corrections and additions to the dictionary. Those are reviewed by the maintainers.

### 2.1 Internal Representation

The internal representation of the dictionary does not contain all inflected forms of each lemma, but rather lists lemmas with a group of special tags describing the lemma from grammatical and lexical standpoints. They are then used to automatically generate the visual representation of the dictionary (see section 2.2 below). Examples of lemmas and corresponding forms are shown in the table 1.

Each tag group starts with the base tag showing the part of speech and the inflection class, if applicable. It may be followed by a series of marks showing specifics of the inflected forms' generation for this lemma, including sound alternations or alternative

endings for certain forms. Additional tags placed after them may indicate supplemental grammatical (e.g. perfective vs imperfective aspect for verbs) as well as semantic information (e.g. animate vs inanimate nouns, specific indication of family names or names of cities). Lastly, the lemma can be marked with flags indicating its usage (colloquialisms, vulgarisms, alternative spellings, orthographic variations, etc.)

деренькотання /n2n
деренькотати /v1.cf.advp :imperf
деренькотіння /n2n
деренькотіти /v1.cf.advp :imperf
деренькучий /adj
дерепресія /n10.p1
дерешуватий /adj
держава /n10.p1.i1k1
державдитор /n20.a.p.ke.< :ua_2019
державець /n22.a.p.<

Table 1: An excerpt from the internal representation of VESUM.

For example, the lemma “деренькотати” (“to jar”) shown in the table 1 is tagged as an imperfective (:imperf) verb of the first conjugation group (v1) that may use synthetic future tense forms (cf) and has a corresponding adverbial participle (advp). Similarly, the lemma “державдитор” (“state auditor”) is described as an animate (<) masculine noun of the second declension group (n20), ending in -a in singular genitive form (a), in -e – in singular vocative form (ke) and having no alternations in plural forms (p). Additionally, it’s indicated to follow the spelling norms introduced by the Ukrainian orthography of 2019 (:ua\_2019).

## 2.2 Generated Visual Representation

The internal representation shown in the previous section provides a way for systematic and economic storage of the lemma descriptions. However, for the dictionary to be used in real-life applications, it is preferable to have a list of all the inflected forms shown explicitly. In the VESUM, one of the forms of such a list is called a visual representation. It is generated from the internal representation and has a set of tags of its own: it copies some semantic and lexical information, removes the tags used solely for mechanical forms generation (e.g. the tags “a” or “ke” shown above to represent certain endings), and adds the characteristics of the inflected forms. Table 2 shows examples

of visual representations for several parts of speech (for the sake of brevity, only part of the inflected forms is shown for each lemma).

As can be seen from the table 2, VESUM generates all the inflected forms for a given lemma, including cases when several alternative versions are possible for a certain form. For instance, the masculine (m) locative (v\_mis) of the lemma “державдитор” (“state auditor”) can surface both as “державдиторові” and as “державдитору”. Possible differences in usage of the alternative forms are indicated as well: the plural (p) accusative (v\_zna) of the adjective “дернуватий” (“soddy”) is “дернуватих” for animates (ranim) and “дернуваті” for inanimates (rinanim).

Verb	Noun	Adjective
деренькотати verb:imperf:inf	державдитор noun:anim:m:v_naz:ua_2019	дернуватий adj:m:v_naz
деренькотать verb:imperf:inf:short	державдитора noun:anim:m:v_rod:ua_2019	дернуватого adj:m:v_rod
деренькочи verb:imperf:impr:s:2	державдиторі noun:anim:m:v_mis:ua_2019	дернуватім adj:m:v_mis
деренькочім verb:imperf:impr:p:1	державдиторові noun:anim:m:v_mis:ua_2019	дернуватому adj:m:v_mis
деренькочем verb:imperf:pres:p:1:subst	державдитору noun:anim:m:v_mis:ua_2019	дернувату adj:f:v_zna
деренькотатиму verb:imperf:futr:s:1	державдиторе noun:anim:m:v_kly:ua_2019	дернуватою adj:f:v_oru
деренькотала verb:imperf:past:f	державдитори noun:anim:p:v_naz:ua_2019	дернуватим adj:n:v_oru
деренькотали verb:imperf:past:p	державдитори noun:anim:p:v_kly:ua_2019	дернуватих adj:p:v_zna:ranim
		дернуваті adj:p:v_zna:rinanim

Table 2: Examples of visual representation in VESUM.

### 2.3 Current Search Form

The search form that has been used for VESUM so far (Rysin & Starcko, 2005–2023; see figure 1) provides only the basic functionality: search across the dictionary’s visual representation by lemmas, inflected forms, or their parts with no options to utilize the grammatical information provided by the dictionary. Apart from that, the results are directly replicating the format of the visual representation, i.e., show only the inflected forms with their tags in the machine- rather than human-readable format.



Figure 1: The previous VESUM search interface.

The limited search capabilities of this form are far from providing users with all the data that can be retrieved from the database. A new user-friendly search interface enabling the creation of search queries based on the grammatical features of the lemmas or their forms would make the database more convenient both for linguistic research and for day-to-day usage as a reference dictionary. Apart from that, structured inflection tables instead of the bare lists of forms would constitute a nice addition to the updated interface.

### 3. Online Dictionaries of Other Languages

The next step in creating the new search interface is an analysis of similar existing tools and surveying the possibility of their adaptation to VESUM. This section provides a short overview of the search engines available for dictionaries and grammatical databases of other languages. It focuses on the Slavic languages since those have similar grammatical categories compared to Ukrainian, and on English as a language with a wide range of lexicographic resources.

### 3.1 Grammatical Dictionary of Polish

The Grammatical Dictionary of Polish (Kieraś & Woliński, 2017; Grammatical Dictionary of Polish; see figure 2) provides a variety of search options for the lemmas and their forms: lexical classes, frequency, gender, aspect, etc. It provides grammatical information, inflection tables, and—for some of the lemmas—clarifications on the meaning. The downside of this dictionary is its unintuitive interface that makes it hard for the users to do an extensive search without studying the documentation for the dictionary.

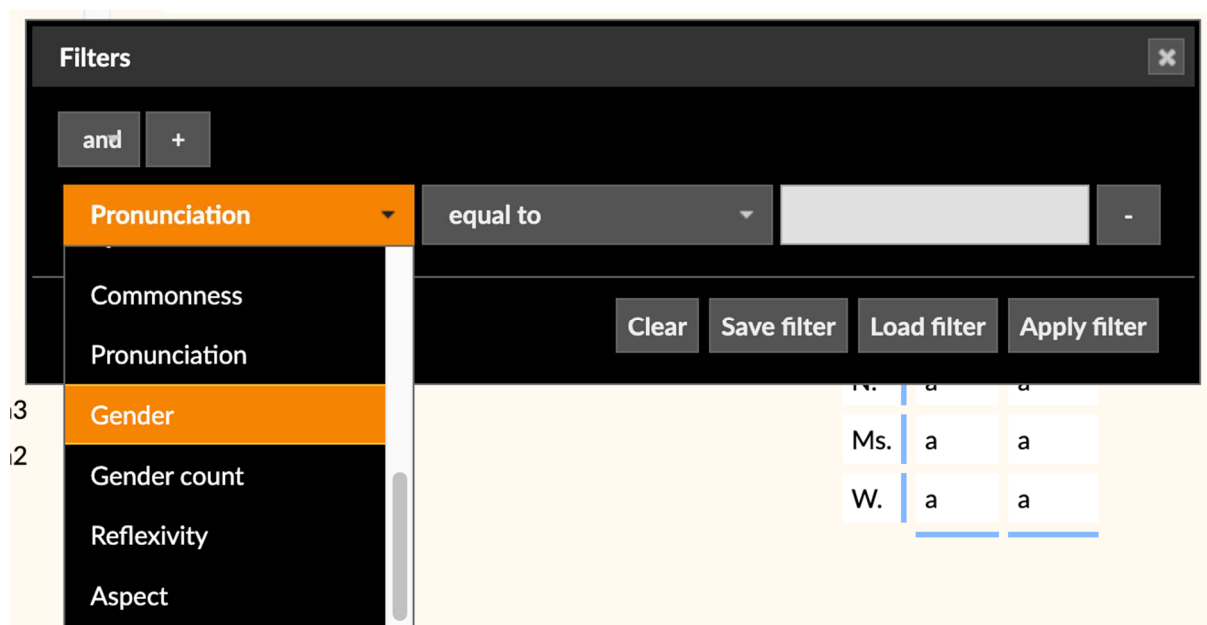


Figure 2: Grammatical Dictionary of Polish.

### 3.2 Grammar Database of Belarusian

The Grammar Database of Belarusian (Bułojčyk & Koščanka, 2021; Grammar Database) also provides a search interface that uses grammatical and lexical information of the lemmas and provides the inflection tables for each inflected lemma. Compared to the Polish resource, it uses a more visual and structured way of filtering by lemma features, which it presents as “search grammars” (see figure 3). Another advantage of this dictionary is the fact that the source code for the search engine is public (Korpus: Corpus Linguistics Software) and is declared to be adjustable for other languages.

Граматыка

Скасаваць Абраць

дзеяслоў

**Пераходнасць**

пераходны  непераходны  пераходны/непераходны

**Трыванне**

закончанае трыванне  незакончанае трыванне

**Зваротнасць**

зваротны  незваротны

**Спражэнне**

1-е спражэнне  2-е спражэнне  рознаспрагальны


Figure 3: Grammar Database of Belarusian.

### 3.3 Dictionaries of English

Some examples of well-known online English dictionaries are the Cambridge Dictionary (Cambridge University Press & Assessment, 2023) and the Macmillan Dictionary (Macmillan Education Limited, 2009–2023). Both provide definitions of the words, their pronunciation, basic grammatical information (for instance, part of speech), and usage examples. In addition, the Cambridge Dictionary supplies some articles with pictures, while the Macmillan Dictionary contains inflection tables and may list synonyms and other related words (see figure 4). Nonetheless, they neither provide the possibility to use any word characteristics in the search queries, nor support search by regular expressions.

The RegEx search for English words however is supported by some other web resources, for example, the Word Finder (Word finder 2023). However, this kind of resource typically doesn't contain any grammatical information as it is mostly oriented at crossword puzzle solving, rather than at providing linguistic information.

**toe** DEFINITIONS AND SYNONYMS ★★

NOUN COUNTABLE UK  /təʊ/

---

**WORD FORMS**

singular **toe**  
 plural **toes**

---

**DEFINITIONS** 2

**1** one of the five individual parts at the end of your foot. Your big toe is the largest, and your little toe is the smallest

*Vera slipped off her shoes and wiggled her toes.*  
*I stubbed my toe (=hurt it by hitting it) on the step.*

**on your toes (=with only your toes on the ground):** *He stood on his toes to look out of the window.*

---

**Synonyms and related words**

General words for limbs and appendages

**ankle** **appendage**

Figure 4: An entry in the Macmillan Dictionary.

## 4. Implementation

Based on the overview of the online dictionaries presented in the previous section, it was reasoned that the Grammar Database of Belarusian has the most fitting search interface for online dictionaries of Slavic languages. One of the most important differences between the VESUM and the Belarusian Grammar Database is the format in which the data is stored. In the VESUM the data is stored in two ways: internal (see section 2.1) with several kinds of tags and marks for each lemma and visual representation (see section 2.2) with a narrower set of tags. The Belarusian Grammar Database utilizes a group of XML files describing paradigms, lemmas, and inflected forms (see table 3). It also contains groups of tags describing each of the items. However, the Grammar Database strongly relies on the tag groups having a certain rigid format and order, so the usage of the suggested format required a transformation of the VESUM tag system. The scripts that achieve that are available on GitHub (Dictionary Format Translator, 2023). The base version of the updated VESUM search uses XML files as its internal data representation. At the time of writing efforts are underway to rebuild the search engine so that it uses the SQL-based data source, in order to achieve a more efficient and seamless operation.



<pre> &lt;Paradigm pdgId="1211000" lemma="па-пя+тае" tag="Z"&gt;   &lt;Variant id="a" lemma="па-пя+тае" pravapis="A1957,A2008"&gt;     &lt;Form tag="" slouniki="krapivabr2012,sbm2012"&gt;па-пя+тае&lt;/Form&gt;   &lt;/Variant&gt; &lt;/Paradigm&gt; </pre>
<pre> &lt;Paradigm pdgId="1127963" lemma="адзіна+ццацера" tag="MAKS"&gt;   &lt;Variant id="a" lemma="адзіна+ццацера" slouniki="piskunou2012:7147" pravapis="A1957,A2008"&gt;     &lt;Form tag="PNP" slouniki="prym2009"&gt;адзіна+ццацера&lt;/Form&gt;     &lt;Form tag="PGP" slouniki="prym2009"&gt;адзінаццацяры+х&lt;/Form&gt;     &lt;Form tag="PDP" slouniki="prym2009"&gt;адзінаццацяры+м&lt;/Form&gt;     &lt;Form tag="PAP" slouniki="prym2009" options="inanim"&gt;адзіна+ццацера&lt;/Form&gt;     &lt;Form tag="PAP" slouniki="prym2009" options="anim"&gt;адзінаццацяры+х&lt;/Form&gt;     &lt;Form tag="PIP" type="nonstandard"&gt;адзінаццацяры+ма&lt;/Form&gt;     &lt;Form tag="PIP" slouniki="prym2009"&gt;адзінаццацяры+мі&lt;/Form&gt;     &lt;Form tag="PLP" slouniki="prym2009"&gt;адзінаццацяры+х&lt;/Form&gt;   &lt;/Variant&gt; &lt;/Paradigm&gt; </pre>

Table 3: Examples of entries in the Belarussian Grammar Database.

Of the two VESUM representation formats, the porting was done for the visual one since it provides the set of data that is closer to the search criteria that the users might be interested in when using the software. Nonetheless, the internal representation does provide some additional data that can be interesting for researchers, so the future development plans include the integration of the internal data into the final search form (see also section 6). The currently supported tags provide information on the part of speech, and POS-specific characteristics for lemmas and their inflected forms: 1) for nouns: animate vs inanimate, common vs proper, abbreviation vs non-abbreviation, gender, number, case, 2) for adjectives: degree of comparison, gender, number, case, usage with animate vs inanimate nouns for certain forms, 3) for verbs: reflexivity, aspect, tense, gender, number, person, 4) for adverbs: degree of comparison, and 5) for conjunctions: coordinating vs subordinating. An example of tag groups describing Ukrainian lemmas is shown in the table 4.

After completing this task, several other items had to be addressed in order to launch the updated VESUM search engine, namely: 1) porting of Korpus search functionality to a more compact framework (from pure Tomcat to Spring Boot), 2) adaptation of search parameters to correspond to the transformed VESUM tag system, 3) adaptation of the inflection tables to correspond to the sets of inflected forms generated by the

VESUM, 4) design adjustments to make the search page more in line with the existing ecosystem of computational linguistic tools for the Ukrainian language. All the listed tasks and future development steps can be followed on the VESUM search GitHub page (2023).

<pre> &lt;Paradigm pdgId="87838" lemma="деренькотати" tag="VMN"&gt;   &lt;Variant id="a" lemma="деренькотати"&gt;     &lt;Form tag="0"&gt;деренькотати&lt;/Form&gt;     &lt;Form tag="R1S"&gt;деренькочу&lt;/Form&gt;     &lt;Form tag="R2S"&gt;деренькочеш&lt;/Form&gt;   [...]     &lt;Form tag="PXP"&gt;деренькотали&lt;/Form&gt;   &lt;/Variant&gt; &lt;/Paradigm&gt; </pre>
<pre> &lt;Paradigm pdgId="87858" lemma="державдитор" tag="NCANM"&gt;   &lt;Variant id="a" lemma="державдитор" orthography="ua_2019"&gt;     &lt;Form tag="NS"&gt;державдитор&lt;/Form&gt;     &lt;Form tag="GS"&gt;державдитора&lt;/Form&gt;     &lt;Form tag="DS"&gt;державдиторові&lt;/Form&gt;     &lt;Form tag="DS"&gt;державдитору&lt;/Form&gt;     &lt;Form tag="AS"&gt;державдитора&lt;/Form&gt;   [...]     &lt;Form tag="VP"&gt;державдитори&lt;/Form&gt;   &lt;/Variant&gt; &lt;/Paradigm&gt; </pre>
<pre> &lt;Paradigm pdgId="88489" lemma="дернуватий" tag="AP"&gt;   &lt;Variant id="a" lemma="дернуватий"&gt;     &lt;Form tag="MNS"&gt;дернуватий&lt;/Form&gt;     &lt;Form tag="MGS"&gt;дернуватого&lt;/Form&gt;   [...]     &lt;Form tag="PAP" options="anim"&gt;дернуватих&lt;/Form&gt;     &lt;Form tag="PAP" options="inanim"&gt;дернуваті&lt;/Form&gt;     &lt;Form tag="PIP"&gt;дернуватими&lt;/Form&gt;     &lt;Form tag="PLP"&gt;дернуватих&lt;/Form&gt;     &lt;Form tag="PVP"&gt;дернуваті&lt;/Form&gt;   &lt;/Variant&gt; &lt;/Paradigm&gt; </pre>

Table 4: Examples of the VESUM entries after the translation of the internal tag system of the VESUM to the one more suitable for the search engine.

## 5. Results

The set-up search system for the VESUM provides users with functionality to perform a search across lemmas or across all the inflected forms in the dictionary, using both

exact queries and regular expressions (see figure 5). The results are displayed as lemmas with lists of their grammatical features. By clicking on a lemma, the user can view its inflection table (see figure 6).

**БЕСУМ**

Слова

Порядок:  Звичайний  Зворотній

Якщо укласти граматику і обрати частину мови, у якій слово може мати декілька форм, можна виводити форму, інакше від початкової

За початковою формою

За всіма формами

Укласти граматику

блакитнавий – прикметник, базова форма  
 блакитненький – прикметник, базова форма  
 блакитний – прикметник, базова форма  
 блакитніти – дієслово, недоконане, незворотнє  
 блакитнішати – дієслово, недоконане, незворотнє

Figure 5: Main page of the developed search interface, search with a regular expression, and a list of displayed results.

одн.	с.	Д.	блакитному
		Зн.	блакитне
		Ор.	блакитним
		М.	блакитнім блакитному
		Кл.	блакитне
	ж.	Н.	блакитна
		Р.	блакитної
		Д.	блакитній
		Зн.	блакитну
		Ор.	блакитною
	М.	блакитній	
	Кл.	блакитна	

Figure 6: A section of an inflection table for the lemma “блакитний” (“light blue”).

The most powerful part of this tool is the search grammars that can be used to filter lemmas by their features. By clicking “Укласти граматику” (“Set up a grammar”) the

user can select a certain part of speech and the POS-specific features of interest. Figure 7 demonstrates a grammar that covers masculine inanimate common nouns.

The screenshot shows a web interface titled "Граматика" (Grammar). At the top right, there are two buttons: "Скасувати" (Cancel) and "Обрати" (Apply). Below the title is a dropdown menu with the text "іменник" (noun). The main content area contains several sections of search criteria, each with a title and a list of options:

- Власна назва** (Proper name):  загальна назва (general name),  власна назва (proper name)
- Істота** (Entity):  істота (entity),  неістота (non-entity)
- Абревіатура** (Abbreviation):  абревіатура (abbreviation),  не абревіатура (not abbreviation)
- Рід** (Gender):  чоловічий рід (masculine),  жіночий рід (feminine),  середній рід (neuter)
- Множинні** (Plural):  множина (plural)
- Відмінок** (Case):  називний відмінок (nominative),  родовий відмінок (genitive),  давальний відмінок (dative),  знахідний відмінок (accusative),  орудний відмінок (instrumental),  місцевий відмінок (locative),  кличний відмінок (vocative)

Figure 7: A search grammar that would capture masculine inanimate common nouns.

## 6. Conclusions and Future Directions

The new search interface that was built for the VESUM is implementing a more user-friendly way of interacting with the database, and by providing advanced search options it allows one to make more use of the information available for each of its items. The use of search grammars (see figure 7) in combination with the regular expressions may be useful for researchers who need to compile lists of words sharing a certain grammatical feature (for example, listing all masculine nouns with typical feminine endings, or verbs that have given prefix, etc.). Apart from that, the neatly structured inflection tables make the dictionary convenient for non-linguists who might be looking for correct spellings of certain words or their forms.

Having said that, there is still a lot of work that can be done on the part of integrating the search form with the VESUM database. The current version implements a search that is based on the tags of the dictionary's visual representation, leaving the internal source files aside. One of the next steps therefore would be to integrate both dictionary representations (as each of them contains some unique information) into the set of search options. That would include grammatical information, such as inflection classes

or alternations, as well as usage mode, such as indications of colloquialisms, vulgarisms, etc.

Other possible directions for improvements include 1) addition of information that might not come directly from the VESUM, e.g., integrating with other online resources for the Ukrainian language, 2) addition of links to the related resources for Ukrainian and other languages, 3) implementing functionality for reporting mistakes and providing suggestions, 4) English localization, and 5) creating the structured public documentation describing all the data available in VESUM and accessible through the search interface.

## 7. References

- Bułowczyk, A. & Koščanka, U. (2021). Belarusian Language Grammar Database. Minsk: Тэхналогія.
- Busel V. (2005). Великий тлумачний словник сучасної української мови: 250000. Kyiv, Irpin: Perun.
- Cambridge University Press & Assessment. (2023). Cambridge Dictionary. Accessed at: <https://dictionary.cambridge.org/>. (26 March 2023).
- Dictionary Format Translator (2023). Accessed at: <https://github.com/tamila-krashtan/dictionary-format-translator>. (26 March 2023).
- Grammar Database. Accessed at: <https://bnkorporus.info/grammar.en.html>. (26 March 2023).
- Grammatical Dictionary of Polish. Accessed at: <http://sgjp.pl/>. (26 March 2023).
- Karpilovska, Y., Kysliuk, L., Klymenko, N. et al. (2013). Активні ресурси сучасної української номінації: Ідеографічний словник нової лексики. Kyiv: TOV “КММ”.
- Kieraś, W. & Woliński, M. (2017). “Grammatical Dictionary of Polish” – an online version. *Jezyk Polski*, 97(1), pp. 84–93.
- Korpus: Corpus Linguistics Software. Accessed at: <https://github.com/alex73/Software-Korpus>. (26 March 2023).
- Krytska, V., Nedozyrn, T., Orlova, L., Puzdyreva, T., Romaniuk, Y. (2011). Граматичний словник української літературної мови. Словозміна: Близько 140 000 слів. Kyiv: Dmytro Burago Publishing House.
- Macmillan Education Limited. (2009–2023). Macmillan Dictionary. Accessed at: <https://www.macmillandictionary.com/>. (26 March 2023).
- Project to generate POS tag dictionary for Ukrainian language. Accessed at: [https://github.com/brown-uk/dict\\_uk](https://github.com/brown-uk/dict_uk). (26 March 2023).
- Rysin, A. & Starko, V. (2005–2023). Large Electronic Dictionary of Ukrainian (VESUM). Web version 6.0.1. Accessed at: <https://r2u.org.ua/vesum/>. (21 March 2023).
- Rysin, A. & Starko, V. (2020). Великий електронний словник української мови (VESUM) як засіб NLP для української мови. *Галактика Слова*. Галині Макарівні Гнатюк, pp. 135–141.
- Shvedova, M. & von Waldenfels, R., Yaryhin, S., Rysin, A., Starko, V., Nikolajenko, T.

et al. (2017-2023): GRAC: General Regionally Annotated Corpus of Ukrainian. Electronic resource: Kyiv, Lviv, Jena. Available at [uacorporus.org](http://uacorporus.org).  
Word finder 2023. Accessed at: <https://findwords.info/>. (26 March 2023).  
VESUM search GitHub page. (2023). Accessed at: <https://github.com/tamilakrashtan/vesum-search>. (26 March 2023).