

VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool

Vasyl Starko¹ and Andriy Rysin²

¹ *Ukrainian Catholic University, 2a Kozelnytska Str., Lviv, 79026, Ukraine*

² *Independent researcher, 104 Hab Tower Pl, Cary, NC, 27513, USA*

Abstract

The paper describes VESUM, a large morphological dictionary of Ukrainian, as a valuable resource for the analysis and synthesis of Ukrainian morphological data. In line with its manifold practical uses, VESUM supplies a rich set of morphological features for more than 400,000 Ukrainian lemmas. Its lexical range extends beyond what is found in Ukrainian monolingual and grammatical dictionaries to cover proper names, abbreviations, alternative spellings, slang, deprecated items, dialectal and archaic words, etc. VESUM's inflectional paradigms include a number of substandard wordforms (marked as such) that occur in texts and need to be recognized by NLP applications. The paper describes VESUM's structure, morphological information it provides, its use in the LanguageTool language checker and in the Lucene search engine, as well as the dynamic tagging component that acts as a complement to the dictionary itself. VESUM's coverage of different text types is also discussed. The dictionary is provided as an open access source via an online repository for the NLP community and is made available online through a web interface in human-readable, searchable format.

Keywords

Morphological dictionary, POS dictionary, Ukrainian, VESUM, POS tagging, morphological analysis.

1. Introduction

Morphology is critical for many downstream NLP tasks, and morphological lexicons have been created for a number of different languages. They are especially useful for highly inflectional languages and are the building blocks of spellcheckers and parsers. The output of a morphological module is often exploited by NLP applications, such as search engines, information extraction systems, and machine translation systems. The high utility and manifold applicability of a large morphological dictionary is convincingly evidenced by such projects as MorfFlex CZ 2.0. Developed stagewise for over 30 years, this Czech lexicon has more than 100,000,000 wordforms representing over 1,000,000 lemmas [5], [18]. Furthermore, grammatical dictionaries are increasingly made available online [23] to present, in contrast to most traditional lexicographical works, inflectional paradigms fully and explicitly. This format is helpful to different groups of users, from non-native students of the language to teachers to professional linguists.

Like other Slavic languages, Ukrainian is highly inflectional: one paradigm may consist of 17-19 forms for a typical noun, 27-32 for verbs (excluding analytical forms), and 32-43 for adjectives. Ukrainian inflectional morphology includes:

- number: singular and plural, with occasionally used, even though generally archaic, forms of dual number;
- gender: masculine, feminine, and neuter;

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

EMAIL: v.starko@ucu.edu.ua (V. Starko); arysin@gmail.com (A. Rysin)

ORCID: 0000-0002-2530-2107 (V. Starko); 0000-0002-8445-3431 (A. Rysin)



© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

- grammatical case: nominative, genitive, dative, accusative, instrumental, locative, and vocative, with a number of words having multiple possible forms for a given case;
- person: 1st, 2nd, and 3rd;
- tense: past, present, and future;
- aspect: imperfective and perfective;
- mood: indicative, imperative, and subjunctive;
- degrees of comparison: positive, comparative, and superlative;

For various reasons, Ukrainian texts abound in spelling and morphological variants and idiosyncrasies, which greatly complicates the task of practical morphological annotation, even when the goal is to handle contemporary texts only. Extending the scope timewise to earlier periods and geographically to Ukrainian diaspora texts necessitates the use of a significantly richer set of morphological devices. There is a distinct need for a large machine-readable dictionary for morphological analysis that would be suitable for various types of texts, able to handle both standard and nonstandard usage, and providing good coverage in Ukrainian corpora. Our goal here is to explain how VESUM meets these challenges, describe the tools and methods underpinning its development, and analyze its practical application.

The paper is organized in the following way. In Section 2, we review the relevant works in the domain of Ukrainian morphology. Section 3 describes the composition and features of VESUM. Section 4 details the applications of VESUM, while the next section provides an account of how POS tagging is performed using VESUM. In Section 6, we discuss text coverage achieved with the help of VESUM. Finally, conclusions are drawn and prospects for future development are outlined.

2. Related works

Ukrainian is a synthetic language in that it expresses grammatical meanings through inflections rather than word combinations. It has a significant degree of irregularity, especially with regard to nominal declension and verbal conjugation. For this reason, an adequate morphological description requires a large number of morphological classes and a number of exceptions to be specified. The creation of a morphological lexicon is, thus, a challenging task, and the situation is compounded by the fact that the existing descriptions of morphological paradigms for Ukrainian do not easily lend themselves to use in NLP. Traditionally, such works single out classes based on both inflections and stress patterns [19] and are, thus, more complicated than is practically necessary for POS tagging, which does not involve accentuation. This approach is adopted in two academic grammatical dictionaries of Ukrainian [9] and [16]. The former comprises 140,000 lemmas (no proper names) and offers an explicit formalized system of morphological codes representing morphological classes, while the latter has over 260,000 lemmas (including proper names) and appears to employ an in-house system of a similar kind. After their initial release in 2011, both resources have not been updated to any substantial degree. Thus, they do not include numerous words that have entered the language over the past decade and an even longer period. Probably due to their academic character, neither dictionary attempts to cover slang, substandard lexical items, abbreviations, and alternative spellings.

In the domain of practical NLP applications, a morphological tagset has been developed and implemented for Ukrainian in an open-sourced project [7], [8] in conformity with a multilingual system that imposes certain restrictions on individual languages. There are also long-running in-house projects [3], [15] that make use of formalized morphological systems.

3. Methods and Materials

As a highly inflectional language, Ukrainian requires a morphological lexicon consisting of lemmas and codes to generate declension and conjugation paradigms, i.e., all the wordforms associated with a given lemma. For a long while, no resource of this type was publicly available for Ukrainian. VESUM [12] was created to fill this gap. In its current version 5.6.0, the dictionary contains over 416,000 lemmas from which over 6.5 million wordforms are generated. VESUM is a

non-commercial project: the dictionary data are available under the CC BY-NC-SA 4.0 license, while its software is distributed under GPLv3.

VESUM has benefited from some of the best lexicographical and morphological resources available for Ukrainian: an academic grammatical dictionary of Ukrainian [9], an academic description of Ukrainian morphology [21], a comprehensive overview of dynamic processes in the modern Ukrainian lexicon [6], an online dictionary collection [13], and other dictionaries.

3.1. Composition of VESUM

As is typical for morphological dictionaries [9], lemmas in VESUM are grouped into inflection classes rather than traditional parts of speech, even though the two largely overlap. For example, the inflection class of verbs does not include participles because they have adjectival declension paradigms. VESUM divides all lexis into 13 inflection classes. Apart from such groupings as nouns (tagged as noun), verbs (verb), adjectives (adj), adverbs (adv), adverbial participles (advp), numerals (numr), conjunctions (conj), prepositions (prep), particles (part), and interjections (intj), VESUM includes onomatopoeic words (onomat), transliterated foreign words (foreign), and non-inflected words (noninfl) that cannot be categorized otherwise.

In the source files, lemmas are grouped into several files, which together comprise the lemma lists. Each entry has the format

```
lemma morphological code stylistic tag comments
```

For example,

```
дружба /n10.p1.ko.< :xp1 # людина
```

```
дружба /n10.p1 :xp2 # стосунки
```

where the slash sign / indicates that the word is inflected; n10 encodes for the specific nominal inflection group; p1 is the code for generating plural forms; ko defines the ending in the vocative case; < is a flag for a person; xp1 and xp2 identify two homonyms; the comments after # specify that the first homonym is a person (groomsman) and the second one refers to a relationship (friendship).

VESUM handles homonymy differently from typical monolingual dictionaries: what it recognizes are grammatical homonyms, i.e., identically spelled words that differ in their grammatical categories (e.g., gender) and/or their paradigms. Homonymous lemmas are marked by indices. In the case of, especially, uninflected words, the syntactic role is taken into account. For example, the entry

```
так adv:&pron:dem|part|conj:coord:subord
```

compactly encodes that this Ukrainian word can function as an adverbial pronoun, a particle, or a conjunction, coordinate or subordinate.

A number of additional tags are conjoined to the basic ones with the ampersand symbol:

&adjp (participle)

&&adjp (also a participle)

&pron (pronoun)

&numr (ordinal numeral)

&insert (parenthetical word/wordform)

&predic (predicative).

Eleven additional tags are used to distinguish the types of pronouns. The complete tagset is available online [12]. More details on the generation of wordforms from lemmas are provided in [17].

Over time, VESUM's range has been extended to cover an increasing number of types that occur in real texts but are outside standard Ukrainian: slang, professional jargon, recent lexical loans, proper names, abbreviations, variant spellings, as well as a reasonable number of dialectal and archaic words. Thus, while VESUM's predominant focus is on morphology, it does supply certain stylistic tags some of which are utilized by the LanguageTool module to warn users about style:

subst (substandard)

coll (colloquial)

arch (archaic, outdated, or, in some cases, dialectal)

slang

vulg (vulgar)

alt (alternative spelling)

ua_1992 (spelling under the 1991 rules)
ua_2019 (spelling under the 2019 rules)
var (variant form)
bad (erroneous or objectionable lemma/wordform)
rare (for items with markedly low frequencies).
Standard neutral vocabulary is left unmarked.

In addition to purely morphological and stylistic tags, VESUM utilizes several semantic tags. These labels are abbr (abbreviation) and prop (proper name) with further specification: prop:lname (last name), prop:fname (first name), prop:pname (patronymic name), prop:geo (geographical name), and prop:abbr (abbreviated proper name).

VESUM's output is a flat list of wordforms in the format
wordform lemma positional tag

For example, here is a fragment of the paradigm for the verb *vesty* (lead):

вести verb:imperf:inf
веди verb:imperf:impr:s:2
ведім verb:imperf:impr:p:1
ведімо verb:imperf:impr:p:1
ведіть verb:imperf:impr:p:2

Positional tags are strings of individual tags each of which encodes a morphological category. The tags are separated by the semicolon: verb, imperfective aspect, imperative mood, singular/plural number, infinitive/person in the example above. Unlike other morphological dictionaries, such as the MorfFlex Dictionary of Czech [18], the names of individual tags are mostly shortened English words. This transparency helps even casual users get a better grasp of the morphological annotation. They can then employ individual tags and their combinations to restrict search queries. This format also conveniently expresses morphological features that can be exploited by rules, taggers, and computer models.

3.2. Features

The distinct features of VESUM that are dictated by its practical orientation and set it apart from other morphological dictionaries can be summarized as follows:

1. Open-source project
2. Machine-readable format
3. Large size (bigger than similar resources)
4. A compact system of inflection codes
5. Dynamic nature (the dictionary is constantly enlarged with new lemmas and is used together with a dynamic tagging component)
6. Wide coverage of proper names: over 54,000 lemmas, including all names of populated areas in Ukraine according to the official register; Ukrainian geographical names introduced in the process of decommunization; more than 3,500 first names, 1,000 patronymic names, and 28,000 last names; a number of foreign proper names
7. Coverage of non-standard vocabulary: 8,000 erroneous lemmas (with replacements), 1,700 most frequent abbreviations, alternative spellings, 1,500 slang words, and over 1,200 archaic words
8. Inclusion of rare morphological forms, such as the colloquial infinitive forms ending in -t' rather than -ty and the variant ending -a for the accusative case of some singular masculine nouns, e.g., *ножа* (knife.Acc.sing)
9. Information on case government, e.g., *rv_oru* after an adjectival lemma means that it governs a noun in the instrumental case
10. Suitability for expansion with other types of linguistic information (phonetic, semantic, etc.) to be applied in the course of text processing.

4. Experiment

One of VESUM's defining features is its practical focus and integration with several related projects. The dictionary is geared toward practical application, handling real-life Ukrainian texts with their complexities and irregularities, formalization of morphological information, and availability in machine- and human-readable form. Its origins can be traced back to the `ispell-uk` project for spellchecking under Linux. The dictionary was later adapted to perform more sophisticated spelling and grammar checking in `Pravopysnyk` [17], the Ukrainian module of the `LanguageTool` language checker [11].

One of the milestones in VESUM's development came in 2017 when a new Ukrainian morphological analyzer was created for the Apache Lucene search engine and the Ukrainian-language Wikipedia articles were re-indexed. Since then, full-text search based VESUM's morphological data has been used also in other web projects. The morphological toolkit containing VESUM has been used in the `lang-uk` project [4] for lemmatization, POS tagging of the `UberText` corpus (over 665 million tokens), and building word vectors. An earlier version of VESUM was converted into the `OpenCorpora` format [2] and used in the morphological module of the `pymorphy2` library and derivative systems for Ukrainian [20].

The most challenging material for VESUM as a dynamic morphological tool has been presented by the General Regionally Annotated Corpus of Ukrainian (GRAC) [14], which is the most diverse corpus of Ukrainian, running a total of over 650 million tokens, spanning over two centuries (1816-2021), composed of over 90,000 texts in various genres written by 20,000 authors who used different spelling systems. VESUM and GRAC form a dynamic tandem: iteratively, VESUM is used to lemmatize and POS tag each new version of GRAC; a list of unrecognized words, sorted by frequency, is then generated from the tagged corpus; new lemmas are extracted by expert linguists from this list, semi-automatically coded, manually verified, and added to VESUM. Subtle modifications have been made in the grammatical apparatus to enable it to handle irregular forms, archaic words, alternative spellings, etc. Over the years, this approach has been mutually beneficial: GRAC has received increasingly better-fitted POS tagging, while VESUM has grown and improved its coverage by drawing lexical items from a wide variety of textual sources.

VESUM is highly sensitive to language change: it includes new geographical names that became official in Ukraine as a result of the 2015 decommunization laws, feminine forms of nouns that have gained currency over the past several years, and morphological and spelling changes introduced in the new official spelling rules of 2019.




While VESUM is primarily intended for machine use, it can also be highly useful to anyone interested in Ukrainian morphology and inflection, to both native speakers and non-native students of Ukrainian. We have developed a web interface for VESUM available at r2u.org.ua/vesum to function as an online grammatical dictionary for a wide audience of human users. Search queries can be adapted (via a checkbox) to focus on lemmas or specific indirect forms. Question marks and asterisks can be used in queries to replace, respectively, one character and zero or more characters. VESUM has evolved over the years to include various forms, such as substandard and archaic forms, that are not normally presented in academic dictionaries of Modern Ukrainian. Figure 1 below illustrates that a typical paradigm for an adjective, such as *зелений* 'green', includes long forms (*зеленая, зеленую, зелене. зеленій*) and the short form *зелен*, which are missing from other such dictionaries. However, they occur in older texts, in elevated speech, in poetry, and in some other instances. Thus, the text-based approach persistently implemented in the compilation of VESUM for an extended period of time leads to enhanced coverage of real language phenomena as compared to other similar resources.

Версія 5.6.0. Налічує 416657 лем ([більше статистики](#), [опис тегів](#))

Докладніше про те, як влаштовано словник, можна дізнатися у цій [статті](#)

Серед усіх словоформ

Знайдено 2 статті

Шукати «зелений» на інших ресурсах:   

```

зелений adj:m:v_naz:compb
зеленого adj:m:v_rod:compb
зеленому adj:m:v_dav:compb
зеленого adj:m:v_zna:ranim:compb
зелений adj:m:v_zna:rinanim:compb
зеленим adj:m:v_oru:compb
зеленим adj:m:v_mis:compb
зеленому adj:m:v_mis:compb
зелений adj:m:v_kly:compb
зелена adj:f:v_naz:compb
зеленая adj:f:v_naz:compb:long
зеленої adj:f:v_rod:compb
зеленій adj:f:v_dav:compb
зелену adj:f:v_zna:compb
зеленую adj:f:v_zna:compb:long
зеленою adj:f:v_oru:compb
зеленій adj:f:v_mis:compb
зелена adj:f:v_kly:compb
зеленая adj:f:v_kly:compb:long
зелене adj:n:v_naz:compb
зеленеє adj:n:v_naz:compb:long
зеленого adj:n:v_rod:compb
зеленому adj:n:v_dav:compb
зелене adj:n:v_zna:compb
зеленеє adj:n:v_zna:compb:long
зеленим adj:n:v_oru:compb
зеленим adj:n:v_mis:compb
зеленому adj:n:v_mis:compb
зелене adj:n:v_kly:compb
зеленеє adj:n:v_kly:compb:long
зелені adj:p:v_naz:compb
зеленії adj:p:v_naz:compb:long
зелених adj:p:v_rod:compb
зеленим adj:p:v_dav:compb
зелених adj:p:v_zna:ranim:compb
зелені adj:p:v_zna:rinanim:compb
зеленії adj:p:v_zna:rinanim:compb:long
зеленими adj:p:v_oru:compb
зелених adj:p:v_mis:compb
зелені adj:p:v_kly:compb
зеленії adj:p:v_kly:compb:long
зелен adj:m:v_naz:short
зелен adj:m:v_zna:rinanim:short

```

Figure 1: The paradigm of the adjective ‘green’ in VESUM’s web interface at r2u.org.ua/vesum.

The web interface also provides quick links that let the user look up the word in question in collection of Russian-Ukrainian and English-Ukrainian dictionaries, in an explanatory dictionary of Ukrainian, and in the GRAC corpus. Moreover, there are links to the full tagset used and statistics.

Other uses of VESUM include the compilation of various types of dictionaries, linguistic research, NLP research, and so on.

5. Results

A static morphological dictionary can hardly be expected to comprehensively cover the multitude of words, such as hyphenated adjectives (for example, *українсько-англійський* ‘Ukrainian-English’) or nominal compounds that are created using active patterns of the language in question. To this end, VESUM has been supplied with a dynamic tagging component that processes lexical items of this kind in Ukrainian texts. Ukrainian has a number of combining forms, such as *бізнес-* ‘business’ and *онлайн-* ‘online’, which are joined with other words with a hyphen, e.g., *онлайн-магазин* ‘online store’. The dynamic tagging component recognizes these and several other types of hyphenated nouns, adjectives, and adverbs with 95% accuracy [17].

VESUM comes together with a set of disambiguation rules that remove a limited number of ambiguous forms based on frequency information and context. The overall task of ambiguity resolution is to be dealt with at a separate stage.

For POS tagging, VESUM is converted into the morfologik format [11] that compactly encodes sequences of the type wordform/lemma/tags. The search is based on finite state automata and is performed using LanguageTool functions. Java code has been added to LanguageTool to carry out dynamic tagging specifically for Ukrainian. The same morfologik format is used in the Ukrainian module of Lucene/ElasticSearch [22], and the content is optimized for search.

For anyone interesting in POS tagging Ukrainian texts using VESUM, the LanguageTool API NLP UK project, available from github.com along with VESUM, provides the TagText utility, along with a tokenizer and a lemmatizer for Ukrainian. TagText calls LanguageTool functions and the Ukrainian dynamic tagging module to perform sentence splitting, lemmatization, POS tagging, and disambiguation. Several hundred cases of ambiguity are also resolved. The user has a choice of receiving TagText’s output in text form or XML and collecting several types of statistical data. The tagged version of Ukrainian text can then be used as input for disambiguation based on transition probabilities and neural networks.

Below is a fragment from the Constitution of Ukraine POS-tagged by TagText with output in xml format:

```
<sentence>
  <tokenReading>
    <token value='Україна' lemma='Україна' tags='noun:inanim:f:v_naz:prop:geo' />
  </tokenReading>
  <tokenReading>
    <token value='е' lemma='бути' tags='verb:imperf:pres:p:1' />
    <token value='е' lemma='бути' tags='verb:imperf:pres:p:2' />
    <token value='е' lemma='бути' tags='verb:imperf:pres:p:3' />
    <token value='е' lemma='бути' tags='verb:imperf:pres:s:1' />
    <token value='е' lemma='бути' tags='verb:imperf:pres:s:2' />
    <token value='е' lemma='бути' tags='verb:imperf:pres:s:3' />
  </tokenReading>
  <tokenReading>
    <token value='суверенна' lemma='суверенний' tags='adj:f:v_kly' />
    <token value='суверенна' lemma='суверенний' tags='adj:f:v_naz' />
  </tokenReading>
  <tokenReading>
    <token value='і' lemma='і' tags='conj:coord' />
    <token value='і' lemma='і' tags='part' />
  </tokenReading>
  <tokenReading>
    <token value='незалежна' lemma='незалежний' tags='adj:f:v_kly:compb' />
    <token value='незалежна' lemma='незалежний' tags='adj:f:v_naz:compb' />
  </tokenReading>
</sentence>
```

Figure 2: Output of the TagText tagger in xml format.

As can be seen, the tagger has not performed disambiguation for these tokens. Only a handful of disambiguation rules are implemented in the tagger, and work continues on full-fledged disambiguation.

6. Discussion

A practical NLP researcher is interested in text coverage, i.e., what percentage of tokens in a given text are recognized by a morphological tagger. To this end, we have generated statistics by using the TagText tagger on Ukrainian texts after filtering out Russian words. The results are presented in Table 1.

Table 1
Text coverage by VESUM

| Corpus | Coverage | | | |
|-------------|------------------|---------------|-------------|---------------|
| | Total word forms | Recognized, % | Total types | Recognized, % |
| science | 18,613,505 | 98% | 674,166 | 76% |
| fiction | 13,713,519 | 97% | 637,052 | 76% |
| news | 19,105,557 | 99% | 577,342 | 85% |
| GRAC sample | 13,409,854 | 98% | 534,047 | 82% |
| wikipedia | 67,373,328 | 95% | 1,973,943 | 47% |

Five corpora (scientific texts, fiction, news, a random GRAC sample, and the Wikipedia corpus from the lang-uk project) have been processed. On non-encyclopedic texts, VESUM achieves a consistent rate of 97-99% in terms of wordforms. The rates of recognized types exhibit greater variance: 76% for scientific texts and fiction; 82% for the GRAC sample, and 85% for news. The Wikipedia corpus [4] is special in that it contains a disproportionately large number of low-frequency proper names. Nevertheless, VESUM achieves text coverage of 95% on Wikipedia articles. An analysis of the unrecognized type list for Wikipedia has revealed the following: unrecognized proper names account for 40% of the total type count; 50% of all unrecognized types have frequencies below 10; 34% occur only once in this corpus; the list contains numerous misspellings, words in Latin script, and alphanumeric expressions.

Bogdan Babych [1] presents data on VESUM's coverage of types, rather than tokens, in four corpora (news, fiction, law, and wikipedia) and develops an algorithm for morphological processing of out-of-vocabulary items. The author reports percentages of unrecognized types that are in agreement with our results for fiction but are higher for news and much more so (by up to 33% percentage points) for the Wikipedia corpus. The higher percentage of unrecognized types in his experiment may be attributed to three factors: 1. The use of an earlier version of VESUM, which has nearly 10% fewer lemmas than the current one. 2. Text quality, such as news scraped from the web, especially with insufficient language filtering that fails to filter out Russian texts. 3. Possible non-use of the dynamic tagging component.

From our experience of processing various large Ukrainian text corpora, the tail of frequency distribution of unrecognized types is composed predominantly of 1) proper names; 2) misspellings and Russian words (written in Cyrillic, which in many cases makes them graphically indistinguishable from Ukrainian words); 3) foreign (mainly English) words written in Latin script. Fiction texts may also contain a number of archaic words and spellings. Given a sufficiently large corpus, especially collected from the Internet, unrecognized types in group 2 may achieve frequencies above 10 or higher. An algorithm for automatic paradigm induction, such as suggested in [1], would greatly increase effective coverage of the first class, generate paradigms for pseudo-lemmas for the second, and have no effect on the third.

VESUM is aimed at handling diverse but legitimate Ukrainian vocabulary. It does cover thousands of lemmas that are outside the standard language, but no attempt is made to cover outright misspellings, Russian vocabulary, or words in Latin script. That said, VESUM's coverage of proper names can be improved in two ways: by adding the more frequent ones to VESUM and by complementing the lexicon with an algorithm along the lines of [1] to treat OOV items. Another area of possible improvement involves coverage of ungrammatical but still common forms. For example, the incorrect form *копозв* (banner.Gen.sing) occurs 11 times in the Wikipedia corpus. It could be added to VESUM and listed alongside the correct form *копозов* with the tag *subst* (substandard). This

way, the incorrect forms will be recognized during POS tagging and supplied with the correct lemma, increasing utility for NLP applications and human users. Many substandard forms have already been incorporated into VESUM, but their inclusion has been somewhat limited because LanguageTool handles these items and misspellings automatically by applying the minimum edit distance algorithm and suggesting correct forms.

7. Conclusions and Further Development

In contrast to other resources for Ukrainian, VESUM is a large dictionary with wider coverage of the Ukrainian word stock, including proper names, abbreviations, non-standard wordforms and lemmas, slang, alternative spellings, and dialect and archaic words. VESUM supplies a series of stylistic and semantic labels, and its efficiency is increased with the help of a dynamic tagging module. VESUM has an accompanying toolkit for the morphological analysis of Ukrainian.

The dynamic tagging module can be enhanced with techniques similar to those suggested in [1]. This approach may address the issue of new terminology and proper names that appear in texts and are not (yet) covered by VESUM.

The morphological dictionary can be complemented with a semantic lexicon to enable both morphological and semantic tagging of Ukrainian texts in one pass.

Overall, VESUM is a powerful open-access source of morphological data for Ukrainian that is already used in several large-scale projects. It achieves high text coverage on various types of texts and can be effectively used in computational linguistics research and NLP applications. It also serves as a rich source of morphological data to a wide range of users via a searchable web interface. With its practical text-oriented approach and growing lemma count, VESUM is a useful and dynamic tool for the evolving Ukrainian language.

8. References

- [1] B. Babych. Unsupervised Induction of Ukrainian Morphological Paradigms for the New Lexicon: Extending Coverage for Named Entities and Neologisms Using Inflection Tables and Unannotated Corpora. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, 2019, pp. 1–11.
- [2] D. Chaplinsky. LT2OpenCorpora. 2022. URL: <https://github.com/dchaplinsky/LT2OpenCorpora>
- [3] N. Darchuk. *Kompiuterna linhvistyka [Computational Linguistics]*, Kyiv University, Kyiv, 2008.
- [4] V. Dyomkin, D. Chaplinsky, A. Stegnii, O. Marikovskyy, V. Tykhonov, O. Petriv, S. Shekhovtsov, M. Chalyi, T. Kodliuk, M. Pavliuchenko, O. Kunikeyevych, Kh. Skopyk. *Lang-uk*. 2022. URL: <https://lang.org.ua/en/corpora/>
- [5] J. Hajič, J. Hlaváčová, M. Mikulová et al. MorFlex CZ 2.0, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. 2020. URL: <http://hdl.handle.net/11234/1-3186>
- [6] N. Klymenko, I. Karpilovska, L. Kysliuk. *Dynamichni protsesy v suchasnomu ukrainskomu leksykonii [Dynamic Processes in the Modern Ukrainian Lexicon]*, Dmytro Burago Publishing House, Kyiv, 2008.
- [7] N. Kotsyba. Overview of the Ukrainian language resources within the multilingual European MULTEXT-East project, v. 4, in: *SISN*, 770, 1 (2013): 122–129.
- [8] N. Kotsyba, I. Shevchenko, I. Derzhanski, A. Mykulyak. *MULTEXTEast Morphosyntactic Specifications, version 4. 3.11. Ukrainian Specifications*. 2010. URL: <http://nl.ijs.si/ME/V4/msd/html/msd-uk.html>.
- [9] V. Krytska, T. Nedozym, L. Orlova, T. Puzdyrieva, Yu. Romaniuk. *Hramatychnyi slovnyk ukrainskoi literaturnoi movy. Slovozmyna [Grammatical Dictionary of the Ukrainian Literary Language. Inflection]*, Dmytro Burago Publishing House, Kyiv, 2011.
- [10] M. Miłkowski. Developing an Open-Source, Rule-Based Proofreading Tool. In: *Software – Practice and Experience*, 40 (7) (2010): 543–566.
- [11] M. Miłkowski, D. Weiss. *MORFOLOGIK*. 2022. URL: <https://github.com/morfologik/morfologik-stemming>

- [12] A. Rysin and V. Starko. Large Electronic Dictionary of Ukrainian (VESUM). Version 5.6.0. 2005–2022. URL: https://github.com/brown-uk/dict_uk
- [13] Russian-Ukrainian Dictionaries. Comp. by A. Rysin, V. Starko, Yu. Marchenko, O. Telemko et al. 2007–2022. URL: <https://r2u.org.ua>
- [14] 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://uacorp.us.org/>
- [15] V. Shyrovok et al. Korpusna linhvistyka [Corpus Linguistics], Dovira, Kyiv, 2005.
- [16] V. Shyrovok et al. “Slovnyky Ukrainy” online [“Dictionaries of Ukraine” online], 2001–2022. URL: <https://lcorp.ulif.org.ua/dictua/>
- [17] V. Starko, A. Rysin. Velykyi elektronnyi slovnyk ukrainskoi movy (VESUM) iak zasib NLP dlia ukrainskoi movy [Large Electronic Dictionary of Ukrainian (VESUM) As an NLP Tool for Ukrainian], in: Halaktyka Slova [Lexical Galaxy], Dmytro Burago Publishing House, Kyiv, 2020, pp. 135–141.
- [18] B. Štěpánková, M. Mikulová, J. Hajič. The MorFFlex Dictionary of Czech as a Source of Linguistic Data, in: Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Democritus University of Thrace, Thrace, Greece, 2020, pp. 387–392.
- [19] O. Taranenko. Slovozmina ukrainskoi movy [Inflection of the Ukrainian Language]. Nyíregyháza, Hungary, 2003.
- [20] N. Tmienova, B. Sus’. System of Intellectual Ukrainian Language Processing, in: ITS 2019 (2019): 199–209.
- [21] I. Vykhoanets, K. Horodenska. Teoretychna morfologhiia ukrainskoi movy [Theoretical Morphology of Ukrainian]. Pulsary, Kyiv, 2004.
- [22] D. Weiss. Ukrainian Morfologik Analyzer. 2022. URL: <https://github.com/apache/lucene/tree/2183756f1c8253002bb697bdb8e026e86c4b3db5/lucene/analysis/morfologik/src/java/org/apache/lucene/analysis/uk>
- [23] M. Woliński, W. Kieraś. The Online Version of Grammatical Dictionary of Polish, in: LREC 2016. Computer Science (2016): 2589–2594.