Corpus-Based Vocabulary Profiling for Ukrainian: From Lexical Analysis to the PULS Digital Learning Platform

Olena Synchak[1], Vasyl Starko[1], Mariana Burak[1], Mykhaylo Svystun[2]

[1] Ukrainian Catholic University, 2a Kozelnytska Str., Lviv, 79026, Ukraine
[2] independent researcher
E-mail: o_synchak@ucu.edu.ua, v.starko@ucu.edu.ua, mburak@ucu.edu.ua, michael.svystun@gmail.com

Abstract

While CEFR-aligned lexical profiles exist for many major languages, Ukrainian as a Foreign Language (UFL) still lacks an empirically grounded vocabulary list. Existing UFL word lists rely largely on the compilers' judgment rather than on systematic corpus data. This study addresses that gap by developing a CEFR-aligned vocabulary profile (A1–C2) through a corpus-based, expert-guided approach. We have compiled a one-million-word corpus from 21 UFL textbooks using Ukrainian NLP tools (NLP-UK toolkit and VESUM dictionary) for lemmatization and tagging. This yielded a list of over 37,000 lemmas, each supplied with frequency and distribution data (across levels and textbooks). Ukrainian general-purpose corpora (GRAC and BRUK) have also been utilized for frequency data extraction. Selected items are categorized by part of speech and thematic group, with CEFR levels assigned using the "significant onset of use" principle. Expert review has proceeded in two stages: external alignment (independent labeling) and internal alignment (refinement via semantic and derivational grouping). The first 5,891 words are published on the PULS platform (puls.peremova.org), a searchable digital resource, filtered by CEFR level, topic, and part of speech. This profile lays the groundwork for the first Ukrainian Learner's Dictionary, which will feature CEFR labels, definitions, corpus examples, illustrations, and audio recordings. The project supports learners, educators, and test developers by establishing a reliable lexical foundation for UFL.

Keywords: learner's dictionary; Ukrainian as a foreign language (UFL); vocabulary profile; learning platform; corpus; CEFR

1. Introduction

Lexical profiles specify lists of high-frequency vocabulary for foreign or second (L2) language learners, distributed across the proficiency levels defined by the Common European Framework of Reference for Languages (CEFR). Such profiles have been developed for several languages, such as English, German, Swedish, French, Estonian, and Slovenian. Researchers emphasize their value in designing effective teaching materials, promoting learner autonomy, particularly through intelligent computer-assisted language learning systems (Volodina et al., 2024), and potentially guiding the development of CEFR-aligned large language models.

Due to their important role in creating a comprehensive learning system (the concept by Barry O'Sullivan, 2021), lexical profiles serve as a bridge between language teaching and proficiency testing (Coxhead, 2011), aligning vocabulary learning objectives with proficiency test constructs.

For Ukrainian as a foreign language (UFL), several efforts have been made to create frequency-based word lists (Borodin & Turkevych, 2023; Buk, 2006a; Partyko, 2004) in response to the clear demand for level-based vocabulary. However, these attempts remain scattered, often lack consistency and methodological rigor, and typically do not exceed 3,000 words. Aligning Ukrainian language education with CEFR standards has become increasingly urgent, especially in the context of developing valid and reliable language proficiency assessments.

CEFR-aligned lexical profiles are typically compiled using either corpus-based frequency data or expert judgment. Recently, a hybrid approach has gained prominence, combining frequency indicators with expert evaluation (Pintard & François, 2020). Such a comprehensive approach appears most suitable for compiling a Ukrainian lexical profile, as most existing UFL resources to date have relied predominantly on one approach at the expense of the other. In particular, the Standard of Ukrainian as a Foreign Language (A1–C2), developed in 2018 and approved in 2024 (SLSUFL), relies solely on professional expertise without a data-driven foundation or detailed vocabulary range specifications.

The development of a Ukrainian Vocabulary Profile is further complicated by the prevalence of level-straddling textbooks, significant variability of vocabulary across learning materials, and the inherent inflectional complexity of the language. We address these issues by using a comprehensive, data-based approach to vocabulary classification. Specifically, corpus-derived data undergo a two-stage CEFR alignment process during expert evaluation, a method particularly effective for morphologically rich languages like Ukrainian.

This article outlines the key challenges in developing a CEFR-aligned vocabulary list for UFL and proposes data-based solutions drawn from corpus analysis and expert review. First, it describes the methodology used to compile the Ukrainian Vocabulary Profile, including corpus design, item selection procedures, expert evaluation, level labelling, and theme assignment across CEFR levels. Special attention is given to discrepancies between corpus-derived frequency data and expert judgments, along with a proposed approach for handling borderline cases. The level assignment was further refined through the semi-automatic grouping of lexical items according to their thematic, semantic, and derivational relations. Finally, the article presents the structure and functionality of the PULS Digital Learning Platform, which integrates proposed solutions into a user-oriented lexicographic and linguodidactic resource.

2. Related Work

2.1 Relevant Lexical Resources for Ukrainian as a Foreign Language

One of the key challenges in creating a lexical profile is the selection of an appropriate corpus as the data source. On one approach, lexical profiles for foreign language learners are developed using quantitative data derived from native language corpora (L1). For example, a recently published basic vocabulary list of 1,000 most frequent words at the A1 level for Ukrainian (Borodin & Turkevych, 2023) is based on data from a general-language corpus of Ukrainian. While useful for beginners, it does not extend coverage to further levels and does not involve L2 data. Earlier, frequency dictionaries, each comprising some 3,000 words extracted from native-speaker texts, were compiled for three functional styles of Ukrainian (Buk, 2006a, 2006b, 2006c). However, they were not designed to meet UFL needs and lacked CEFR-level indicators.

Contemporary researchers caution against directly applying frequency data from general-language corpora to vocabulary lists for foreign language learning (Alfter, 2021). Such corpora reflect native speaker usage, which often diverges from the communicative needs and content of L2 materials, and typically lack explicit mappings to proficiency levels (Pintard & François, 2020). Therefore, the applicability of the abovementioned resources to UFL needs may be limited, and their data would need to be carefully validated in any case.

Zinoviy Partyko's (2004) core vocabulary list was an innovative resource for its time, combining frequency and variance data from four UFL textbooks and frequency dictionaries of fiction and journalism, including L1 learner texts. Using a ‘concentric circles’ approach, it grouped words by shared presence across sources and organized them into three proficiency levels. However, these levels are not aligned with the CEFR, and the resource conflates data for receptive and productive skills, as well as L2 and L1 learner corpora.

Since L1 corpora are generally unsuitable for developing L2 frequency lists, we argue that texts specifically designed for L2 learners should serve as the primary source for building a corpus aimed at generating a frequency list for the UFL.

2.2 Approaches to Compiling CEFR-Aligned Lexical Profiles

The development of the Ukrainian Vocabulary Profile aligns with the broader practice of creating Reference Level Descriptions (RLDs) that operationalize CEFR descriptors through empirically grounded inventories of lexical, morpho-syntactic, and discourse features (Green, 2010). While the newly approved UFL Standard (2024) outlines RLDs for Ukrainian, it lacks lexical specification. The Ukrainian Vocabulary Profile addresses this gap by providing level-specific vocabulary ranges.

Lexical profiling is now commonly based on L2 corpora: textbook corpora inform receptive skills, while learner corpora are used for productive skills (Alfter, 2021). For instance, the English Profile relies on learner data, reflecting productive use (Capel, 2010), whereas the CEFRLex project created receptive word lists for five languages. The Swedish Profile uniquely includes both receptive and productive lists (Volodina et al., 2024). The focus thus depends on the intended use of the word list.

Recent lexical profiling approaches consider both word frequency and dispersion within a corpus (Alfter et al., 2016). A key issue remains the criteria for assigning words to proficiency levels—whether based on their first appearance in a textbook (the "First occ" approach) or on frequency at a given level (Pintard & François, 2020). The **First occ** approach, used in the CEFRLex interface as well as in Swedish and Estonian profiles, has proven effective for automating level assignment and predicting levels for new words (Pintard & François, 2020). However, it may introduce bias, since texts aimed at developing receptive skills are typically constructed to include some unfamiliar vocabulary to support the development of learners' compensatory strategies.

Conversely, Alfter et al. (2016) proposed the **"significant onset of use**" approach, which refines level assignment by identifying the first CEFR level where a word shows a marked increase in use across learners. Rather than relying on raw frequency, this method considers word diversity, namely usage frequency relative to the number of learners, ensuring more reliable, learner-centered assignments (Alfter et al., 2016). Initially applied to Swedish learner essays, it also proved effective in developing a Slovenian textbook-based frequency list (Klemen et al., 2023). In our project, we use the idea of the "significant onset of use" but apply it differently. Instead of measuring word diversity as proposed in the original study (Alfter et al., 2016), we rely on expert evaluation that considers quantitative data that is available to us, namely frequency and dispersion across textbooks by CEFR level.

Furthermore, researchers emphasize the importance of grouping words to aid vocabulary acquisition, even though opinions vary on how such grouping should be approached. On the one hand, Nation (2000) argued against teaching vocabulary strictly in lexical sets, noting that words within the same thematic group often vary greatly in frequency and educational relevance. Drawing on several studies from the late 1990s, he showed that presenting synonyms and similar words together can cause interference and hinder learning. Instead, he advocates spacing word introduction over time to ensure stable mastery and minimize confusion, emphasizing a balance between frequency-based selection and the need to minimize interference while also prioritizing the natural language use (Nation, 2000).

On the other hand, researchers emphasize that teaching Slavic vocabulary to foreign learners benefits from focusing on word or derivational families (see Pastuchowa, 2009, for Polish). Introducing learners to word-formation mechanisms enhances vocabulary acquisition and morphological awareness, enabling them to understand and create new words from common roots (Kardela, 2015). Thus, word productivity is increasingly recognized as a key criterion in frequency list selection (Kaczmarek, 2006). In this perspective, corpus-based frequency data should be supplemented with measures of derivational productivity, as words with greater capacity to generate new items hold higher pedagogical value.

In our view, the two opinions outlined above are not contradictory but mutually complementary. Specifically, we propose grouping words into derivational families within the lexical profile. At the same time, following Nation's recommendations, we distribute related words across different proficiency levels, taking care to minimize potential interference and support steady lexical progression.

Since the UFL learner corpus is still under development and lacks sufficient data for reliable level assignment, our receptive vocabulary list is based on L2 textbooks. We compiled a one-million-word corpus from 21 textbooks covering CEFR levels A1 to C2. This approach supports the initiative of the Ukrainian Catholic University (UCU) to develop CEFR-aligned proficiency tests for UFL (SOUL Test) and ensures the lexical profile's relevance for receptive skills assessment. Word frequency and dispersion in the corpus serve as primary inputs for expert-guided level assignment, which is further refined through semi-automatic thematic, semantic, and derivational grouping.

3. Methodology

The methodology of this project combines corpus building, frequency word list generation with professional expertise and lexical analysis.

3.1 Source Corpora

Corpora are a must-have resource for virtually any contemporary lexicographic project. The Ukrainian Vocabulary Profile described here rests on a solid corpus foundation. More specifically, three corpora are envisaged for the entire project:

1. A UFL Textbook Corpus spanning A1-C2 levels as the main source of linguistic data for receptive skills.

2. A balanced Corpus of Modern Ukrainian will enable close examination of the selected vocabulary in context, providing frequency data, collocations, usage examples, etc. for the learner's dictionary. With the projected size of 100 million words, it will be more than sufficient for the tasks. Critically, it is designed to include a significant proportion of spoken Ukrainian, which is not found in most Ukrainian corpora.

3. A Ukrainian Learner Corpus will serve as the source of productive language data with respect to learners of UFL.

Despite a growing number of Ukrainian corpora, none of the above types was available at the beginning of the project. Thus, a decision was made to build these types of corpora in the order mentioned above. Currently, the **textbook corpus** has been constructed and is ready for use (see section 4.1 below), while the other two corpora are under development.

Texts specifically designed for L2 learners offer the most reliable foundation for generating frequency lists tailored to the needs of foreign learners. At the same time, owing to discrepancies in vocabulary distribution across CEFR levels in UFL materials, we have decided to cross-check these data with frequency information from general-language corpora. This ensures that the selected words are truly relevant for comprehending texts created not only for educational purposes but also by native Ukrainian speakers.

To this end, we used data from **two general-language corpora**: (1) the Ukrainian Brown Corpus (BRUK), the only balanced Ukrainian corpus available to us, and (2) a subcorpus of the larger and more versatile General Regionally Annotated Corpus of Ukrainian (GRAC) (Shvedova et al., 2017–2025; Shvedova, 2020). While BRUK is useful for high-frequency lemmas, and GRAC is informative even for less frequent vocabulary, the UFL textbook corpus best captures curriculum-based vocabulary. For level assignment, the frequency-ranked lemma list from the textbook corpus served as the primary guide, while frequency ranks from general-language corpora played a supplementary role, providing additional reference points.

3.2 Lexical Inventory: Procedure and Structure

Various approaches exist for compiling word lists, using lemmas, word forms, derivational families, or a combination of these as a basis (Nation & Waring, 1997). Our word lists were initially created based on lemmas. Lemma frequency lists were automatically generated for each textbook and the BRUK corpus using the TagText tagger (TagText), while the corresponding data was extracted from the GRAC subcorpus through its corpus manager. For convenience, the resulting lists were merged into three spreadsheets (A1-A2, B1-B2, and C1-C2). The experts processed lemmas following their frequency ranks in the textbook corpus, while simultaneously considering their absolute frequencies and distribution in this corpus, as well as their ranks (and to a lesser degree, absolute frequencies) in the two general-language corpora, BRUK and GRAC.

Later, lemma frequency lists were manually reviewed and expanded to include relevant word forms, phrases, and, in rare cases, distinct senses. To enhance the accuracy and pedagogical usability of the wordlist, several post-processing steps were applied. First, phonetic variants (e.g., *у/в* ‘in’, *і/й* ‘and’) were grouped under a single entry. Second, high-frequency formulaic expressions such as *будь ласка* 'please', *Добрий день!* 'Good afternoon!', and *Як справи?* 'How are you?' were added as separate items. Third, for clarity at the beginner level, imperative forms of classroom verbs (e.g., *Пишіть!* 'Write!', *Повторюйте!* 'Repeat!') were listed separately from their lemmas to reflect limited paradigm exposure. Fourth, gendered pairs (e.g., *турист/туристка* 'tourist') and aspectual verb counterparts (e.g., *видавати/видати* 'to publish') were added to fill lexical gaps. Fifth, polysemous words were manually disambiguated where necessary (*дорогий* 'expensive' vs. 'dear'; *голова* 'head' [body part] vs. 'head' [person]). However, systematic word sense disambiguation will be addressed in the next phase, during the development of the CEFR-labeled Ukrainian learner's dictionary. Finally, proper names were tagged by topic but not assigned CEFR levels.

After words were assigned to a specific level (see 3.3 and 3.4 for more details), they underwent thematic categorization. The list of themes was compiled based on the analysis of several UFL sources, such as the thematic lists in the UFL Standard and Yabluko textbooks (Burak, 2015; Synchak, 2015; Bartkiv & Borodin, 2022). The identified themes were compared with those in the English Vocabulary Profile (EVP), the Cambridge B1 Preliminary Vocabulary List (2020), and the thematic specifications for the Breakthrough level, developed for the English Profile in consultation with ALTE (Breakthrough, 2009). Based on this analysis, a classification scheme of 39 topics was established (see Table 1).

Themes	Themes
1 House, apartment, home	21 Descriptions of things
2 Grammatical terms	22 Personal feelings, thoughts and experiences
3 Actions	
4 Etiquette formulas	23 Weather
5 Hobbies and leisure	24 Politics
6 Crime	25 Shopping
7 Names	26 Services
8 Food and drinks	27 Nature
9 Colors	28 Jobs, professions
10 Communication	29 Plants
11 Countries, nationalities, languages	30 Holidays and celebrations
12 People: appearance	31 Sports
13 People: character	32 Relationships
14 Furniture and interior	33 Animals
15 Art and entertainment	34 Technology and communications
16 City: objects and places	35 Body and health
17 City: institutions, buildings	36 Transportation, travel, traveling
18 Training and education	37 Finance
19 Names of cities	38 Time
20 Clothing, footwear and accessories	39 Numbers

Table 1: Thematic Classification Scheme for the PULS Platform

3.3 Expert Evaluation

The CEFR alignment of the Ukrainian Vocabulary Profile was carried out through a two-stage expert evaluation process: external alignment and internal alignment (Figure 1), in line with recommendations from *Aligning Language Education With the CEFR: A Handbook* (2022).

At the external alignment stage, two experts, both experienced UFL instructors and authors of UFL textbooks, independently assigned CEFR levels to lexical items. One expert also has extensive experience teaching English, and the other has compiled a corpus-based dictionary. Before the evaluation, both underwent a familiarization process and specialized training to ensure a shared understanding of CEFR levels and descriptors. Word frequency, distribution across textbook levels, and corpus frequency data served as the primary reference points in their assessment.

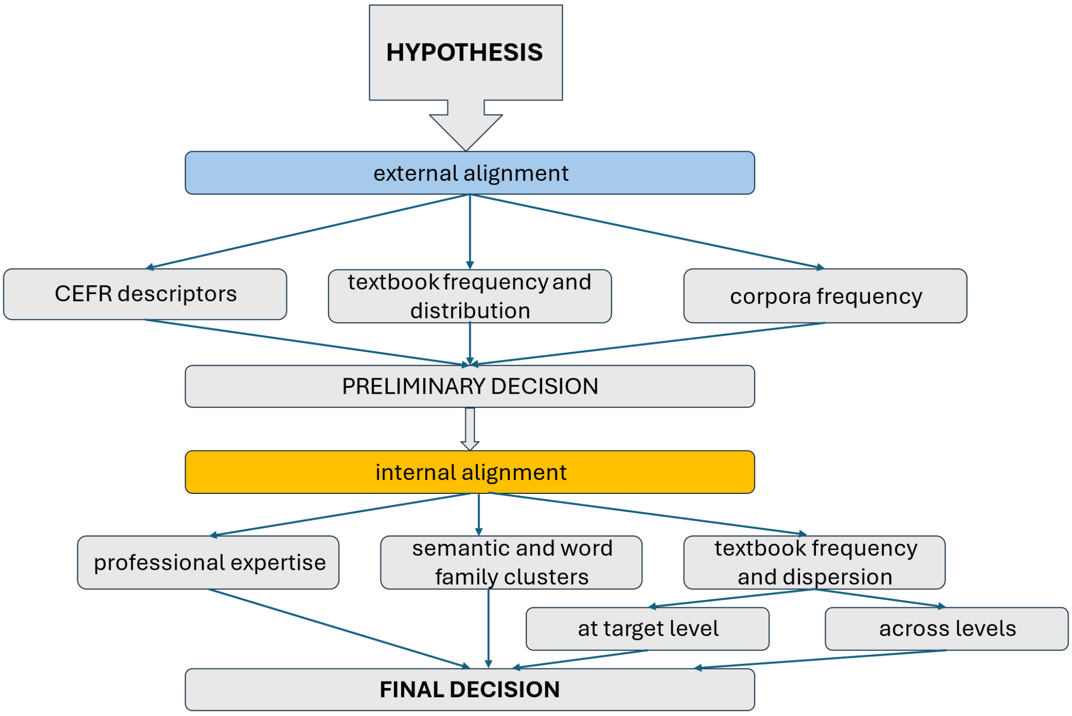


Figure 1: Expert Decision-Making Process

During the internal alignment and standardization process, lexical items underwent semantic and derivational analysis and were semi-automatically organized into thematic, semantic, and word-family sets to maintain coherence and consistency across levels. This allowed for adjustments to the initial assignments and the resolution of expert disagreements, drawing on further analysis of usage examples from both textbooks and general-language corpora. Grouping also aids level verification during the validation phase, which serves as an ongoing process of quality control.

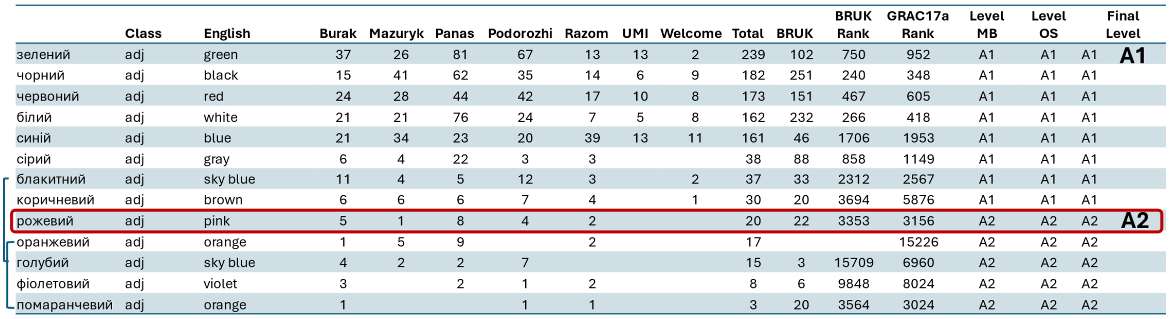
3.4 Level Assignment

In assigning CEFR levels, we addressed concerns about the limitations of labelling words in isolation. Following Pintard and François (2020), who highlight the role of lexical networks in L2 acquisition, we complemented individual-word annotation with thematic, semantic (e.g., synonyms and antonyms), and derivational grouping. This shift toward lexical networks was especially important during the standardization phase, where expert judgments were aligned.

First, each expert independently assigned CEFR levels to the words on the list. The evaluation was not always conducted sequentially; rather, the expert may organize words into thematic or semantic groups to ensure logical progression. A good illustration of this process is the classification of color names (see Table 2 below).

3.4.1 Example of External Level-Labelling Alignment

Relying on frequency and distribution data in A1-A2 level textbooks, color names with high frequency in both textbooks (30–239 occurrences) and general corpora that appear in 6 or 7 beginner-level textbooks, were assigned to the A1 level. In contrast, color names with a lower total absolute frequency (<20) found in 3–5 textbooks were classified at the A2 level. As a result of external level-labelling alignment, the words *блакитний* and *голубий*, both referring to 'sky blue', were assigned to A1 and A2, respectively. In contrast, *оранжевий* and *помаранчевий*, both meaning 'orange,' were assigned to the same level, A2. However, these assignments will be reviewed at the next stage of internal level-labelling alignment.

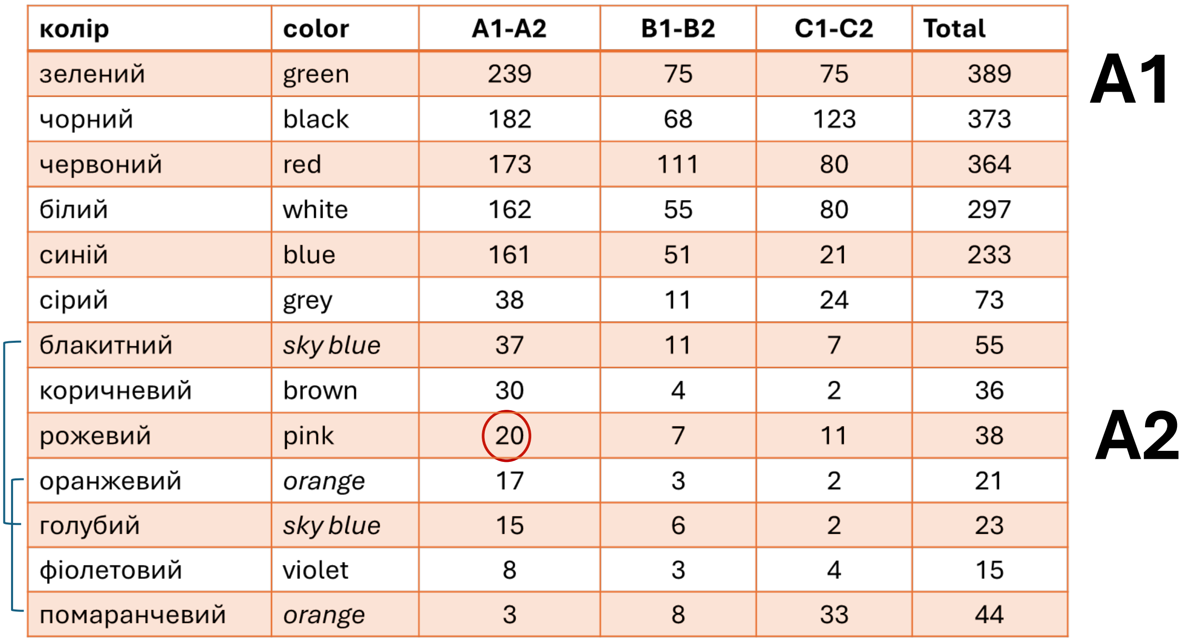


												BRUK	GRAC17a	Level	Level		Final
	Class	English	Burak	Mazuryk	Panas	Podorozhi	Razom	UMI	Welcome	Total	BRUK	Rank	Rank	MB	OS		Level
зелений	adj	green	37	26	81	67	13	13	2	239	102	750	952	A1	A1	A1	**A1**
чорний	adj	black	15	41	62	35	14	6	9	182	251	240	348	A1	A1	A1	
червоний	adj	red	24	28	44	42	17	10	8	173	151	467	605	A1	A1	A1	
білий	adj	white	21	21	76	24	7	5	8	162	232	266	418	A1	A1	A1	
синій	adj	blue	21	34	23	20	39	13	11	161	46	1706	1953	A1	A1	A1	
сірий	adj	gray	6	4	22	3	3			38	88	858	1149	A1	A1	A1	
блакитний	adj	sky blue	11	4	5	12	3		2	37	33	2312	2567	A1	A1	A1	
коричневий	adj	brown	6	6	6	7	4		1	30	20	3694	5876	A1	A1	A1	
рожевий	adj	pink	5	1	8	4	2			20	22	3353	3156	A2	A2	A2	**A2**
оранжевий	adj	orange	1	5	9		2			17			15226	A2	A2	A2	
голубий	adj	sky blue	4	2	2	7				15	3	15709	6960	A2	A2	A2	
фіолетовий	adj	violet	3		2	1	2			8	6	9848	8024	A2	A2	A2	
помаранчевий	adj	orange	1			1	1			3	20	3564	3024	A2	A2	A2	

Table 2: Color Name Frequency and Dispersion Across A1-A2

3.4.2 Example of Internal Level-Labelling Alignment

During the internal level-labelling alignment, we considered both the frequency and the dispersion of lexical items across textbooks of different CEFR levels (see Table 3). While experts assigned levels to isolated words at the previous stage, in this phase an individual curator organized the words into thematic groups to determine their final level assignments. As shown in Table 3, color names appear most frequently in A1–A2 textbooks, indicating that they should be allocated to one of these levels. The division between A1 and A2 was made according to the following principle: color terms with frequencies ranging from 239 occurrences (*зелений* 'green') to 30 (*коричневий* 'brown') were assigned to A1, while those with frequencies between 20 and 3 were placed at A2. At this stage, we also revised the earlier distribution of synonyms for orange (*оранжевий* and *помаранчевий*), both of which had initially been placed at A2. To avoid potential lexical interference, as discussed by Nation (2000), and support clarity in early vocabulary acquisition, *оранжевий* was moved to A1, despite frequency data that would otherwise support its placement at A2.



колір	color	A1-A2	B1-B2	C1-C2	Total
зелений	green	239	75	75	389
чорний	black	182	68	123	373
червоний	red	173	111	80	364
білий	white	162	55	80	297
синій	blue	161	51	21	233
сірий	grey	38	11	24	73
блакитний	*sky blue*	37	11	7	55
коричневий	brown	30	4	2	36
рожевий	pink	20	7	11	38
оранжевий	*orange*	17	3	2	21
голубий	*sky blue*	15	6	2	23
фіолетовий	violet	8	3	4	15
помаранчевий	*orange*	3	8	33	44

Table 3: Color Name Distribution Across CEFR Levels

During the internal level-labelling alignment, synonyms are grouped into sets of semantically related items, as shown in Table 4. This case illustrates how partial synonyms, alongside their derivational relations, are distributed across CEFR levels, whereby *традиційний* is assigned to A1, *типовий* to A2, *звичний* to B1, and *стандартний* to B2. Importantly, derived words are not necessarily graded immediately after their base forms. However, the principle tends to be upheld in this particular case: *традиція* and *традиційний* are assigned to A1; *стандарт* is placed at A2, while *стандартний* is classified as B2.

Lemma	English equivalent	C1-C2	B1-B2	A1-A2	Total	Final level
традиція	tradition	120	56	60	236	A1
традиційний	traditional	63	39	54	156	A1
типовий	typical	16	6	17	39	A2
звичний	habitual	31	2	1	34	B1
стандартний	standard *adj*	11	3	3	17	B2
стандарт	standard *n*	11	1	3	15	A2

Table 4: Level Allocation for Words within a Synonym Cluster and their Derivational Relations

Our approach to level assignment can be summarized as follows:

- Basic words are assigned to levels based on their frequency and dispersion in the textbook corpus, as well as frequency data and ranking in general-language corpora.
- Final level assignment was made by an individual curator within thematic and semantic word groups.
- Words from the same thematic group are not necessarily placed at the same level; instead, they are distributed across levels when justified by the above indicators.
- Synonyms are allocated to different levels based on frequency data and/or didactic criteria, ensuring that no level is overloaded with similar items.

3.5 Theme Assignment, Word Families, and Level Verification

3.5.1 Thematic Assignment and Vocabulary Supplementation

After finalizing level assignments and thematic classification, the words were aligned by CEFR level within each of the 39 thematic groups. For instance, the resulting list for the topic “Furniture and interior” at A1 level includes such words as: *стіл* ‘table’, *стілець* ‘chair’, *шафа* ‘wardrobe’, *ліжко* ‘bed’, *лампа* ‘lamp’, *годинник* ‘clock’, *диван* ‘sofa’, *рушник* ‘towel’, *посуд* ‘dishes’, *холодильник* ‘refrigerator’, *меблі* ‘furniture’, *крісло* ‘chair’, *килим* ‘carpet’, *ваза* ‘vase’.

Grouping words into thematic categories helped verify level labelling and identify underrepresented vocabulary in the textbooks. This process revealed gaps in topics such as fish names, body parts, and internal organs, which were subsequently supplemented. Fish names were assigned to two thematic groups (“Food and drink” and “Animals”) and supplied with the guide word “Fish.”

We first analyzed the fish names found in UFL textbooks and frequency of their usage across CEFR levels (Table 5). At A1–A2, only *тунець* ‘tuna’ and *лосось* ‘salmon’ appear; at B1–B2, *оселедець* ‘herring’, *сом* ‘catfish’, and *лосось* ‘salmon’; and at C1–C2, seven names, including *окунь* ‘perch’, *щука* ‘pike’, and *скумбрія* ‘mackerel,’ etc. By placing most fish names at higher levels, this distribution lacks systematic grounding in real-life Ukrainian contexts and limits early learners' access to everyday texts, such as menus or cultural references.

Lemma	English equivalent	C1-C2	B1-B2	A1-A2	Total
оселедець	herring	8	17	0	25
сом	catfish	3	2	0	5
тунець	tuna	1	1	3	5
лосось	salmon	1	1	1	3
окунь	perch	1	0	0	1
щука	pike	1	0	0	1
скумбрія	mackerel	1	0	0	1

Table 5: Fish Names and their Overall Frequency in UFL Textbooks

Therefore, we supplemented this list with additional fish names (Table 6) commonly found in Ukrainian waters and served in food outlets, assigning them to CEFR levels based on cultural context and frequency data from the GRAC corpus.

A2	B1	B2	C1
оселедець ‘herring’ *лосось ‘salmon’*	*форель ‘trout’* *тунець ‘tuna’* *скумбрія ‘mackerel’* *дорадо ‘dorado’* *окунь ‘perch’*	*хек ‘hake’* *карась ‘crucian’* *короп ‘carp’* *щука ‘pike’* *камбала ‘flat-fish’* *сом ‘catfish’*	*кілька ‘sprat’* *вугор ‘eel’* *діал. пструг = форель ‘trout’* *сьомга = лосось ‘salmon’*

Table 6: Fish Names and Their Distribution across CEFR Levels

As with level labelling, thematic grouping was carried out according to CEFR guidelines and pedagogical principles. However, semantic clarity was often prioritized over didactic function to maintain formal consistency within the lexicographic platform. For instance, in assigning a topic to a word *зручний* ‘comfortable,’ a choice had to be made between “Description of things” or “Clothing, footwear, and accessories.” The consistency criterion led to its placement in “Description of Things,” alongside adjectives like *великий* ‘big’, which can describe clothing but also various other items.

In assigning words to thematic groups, we followed several key principles:

- Semantic relevance to ensure accurate formalization;
- Alignment with CEFR recommendations based on linguistic and didactic criteria;
- Consistency across similar cases;
- Allowing words to belong to multiple thematic groups when appropriate;
- Excluding non-thematic items, particularly function words.

3.5.2 Word Families and Level Verification

In addition to thematic classification, grouping words with a common root into derivational families also supported the verification of level assignments. Word families allow grouping derivationally related forms around their base. For example, the items in the word family shown in Table 7 share the root of the base form *писати* ‘to write’ and are assigned CEFR levels based on the principle of significant onset of use.

Lemma	English equivalent	Final level	Word family
писати	write	A1	писати
написати	write down	A1	писати
підпис	signature	A1	писати
дописати	fill in	A2	писати
переписати	rewrite	A2	писати
підписати	sign	A2	писати
записка	note *n*	B1	писати
допис	post n	B1	писати
писатися	spell	B2	писати

Table 7: Level Verification Within the Word Family *писати* ‘to write’

Grouping items into word families helps ensure clearer lexical progression and more consistent labeling across levels. Following the initial expert review, the word family *жити* ‘to live’ was analyzed, as shown in Table 8. It was discovered that two words were placed at A1, five at B1, and none at A2. Further analysis showed that the word *житель* '(male) resident,' while relatively rare in the A1-A2 textbooks (9 occurrences) ranked fairly highly in the frequency lists generated from the general-language corpora, BRUK and GRAC (1,204 and 2,347, respectively). Thus, it was assigned to the A2 level, together with its feminine counterpart *жителька* '(female) resident.'

Lemma	English equivalent	Final level	Word family
жити	live	A1	жити
життя	life	A1	жити
житель	inhabitant	A2	жити
жителька	(female) inhabitant	A2	жити
прожити	live	B1	жити
проживання	residence	B1	жити
проживати	reside	B1	жити

Table 8: Final Level Assignments within the Word Family *жити* 'to live'

This case highlights the limitations of using the significant onset of use approach alone for level assignment. Due to slight corpus imbalance (Table 9) and the presence of lower-level vocabulary in higher-level textbooks, absolute frequencies at C1–C2 can be misleading. For instance, although the words in Table 8 peak at C1–C2, this does not warrant assigning them to these levels. Combining quantitative data with two-stage expert evaluation yields a more reliable level classification.

In general, thematic and word-formation grouping proved useful for reconciling expert evaluations during the final level assignment and for verifying results. In complex cases of disagreement between corpus data and expert judgments, or between experts, additional experts may be consulted in future standard-setting.

4 Results

4.1 Corpus Results

The UFL textbook corpus we have constructed comprises the texts of 21 textbooks, carefully selected by experts with the view to include high-quality, widely used editions and achieve a reasonable balance across the A1–C2 levels. Expert judgement was exercised for level assignment in cases, quite common for UFL, when textbooks straddled levels, e.g., A2–B1 (as declared by the authors). The composition of the textbook corpus is shown in Table 9 below.

	CEFR levels		
	A1–A2	B1–B2	C1–C2
Q-ty of textbooks	7	6	8
Lemmas	9,342	16,602	27,178
Tokens (appr.)	393,000	373,000	539,000
Proportion of total	30%	29%	41%
Total corpus	1,305,000 tokens, 922,000 words, 37,087 unique lemmas		

Table 9: Composition of the UFL Textbook Corpus

A nearly perfect equilibrium has been achieved between A1–A2 and B1–B2 levels (30% and 29% of the total corpus size, respectively), while C1–C2 textbooks justifiably comprise a larger proportion due to the increasing diversity and complexity of the vocabulary at these top levels. The textbook corpus contains approximately 922,000 words and 1,305,000 tokens. It was automatically lemmatized, tagged morphologically, and disambiguated using the TagText tagger for Ukrainian (NLP-UK) and the VESUM dictionary (VESUM). This tagger regularly achieves over 99% accuracy for lemmatization in Modern Ukrainian texts, which is crucial for the downstream tasks of vocabulary processing. Additionally, manual verification was carried out in the UFL textbook corpus to fix any incorrect lemmatizations, resulting in nearly 100% accuracy. The UFL textbook corpus (UFLTC), complete with the lemma, part of speech, and full morphological specification for each word, has been made available online for internal use by the researchers via the NoSketchEngine corpus management system.

While the UFL corpus contains some 37,000 lemmas, which is more than three times the number necessary for the Ukrainian Vocabulary Profile, 37% of these are hapax legomena; 32% occur 2–5 times, and 10% are recorded 6–10 times. This kind of scatter places more weight on the cross-check with frequency indicators from general-language corpora, especially at higher levels, and further underscores the need for the Ukrainian Vocabulary Profile in order to enhance consistency across future UFL textbooks and other materials.

While our 100-million-word balanced corpus of Ukrainian is still under development, the experts have relied on frequency lists generated from two general-language corpora, BRUK and GRAC. Distinguished by its carefully balanced composition, high-quality written texts, and meticulous annotation and disambiguation (Starko & Rysin, 2023), the BRUK corpus has been utilized for cross-checking only the most frequent vocabulary as warranted by its small size. We have also extracted frequency information from a GRAC subcorpus using the following parameters through the

NoSketchEngine interface: corpus version 17a (the only disambiguated version available), fiction and spoken texts, time period 2000–2023, and 10 million random concordance lines. The resulting frequency list contains more than 120,000 words and has sufficient depth for our purposes. For example, the word *алгоритм* ‘algorithm’ is ranked 11,274 with the absolute frequency 50, which means that the list contains plenty of data beyond the top-ranked 10,000 words.

4.2 Expert Evaluation and Lexical Distribution Challenges

The expert evaluation process combined corpus-based frequency data with professional linguodidactic judgment, structured around three main thresholds of frequency in the textbook corpus. For the ~1700 most frequent items in the A1–A2 textbook subcorpus (9,822 to 17 occurrences), level assignment was largely straightforward. These words were labeled based on three converging indicators:

1) frequency in the UFL textbook corpus,

2) distribution across A1–A2 textbooks, and

3) frequency ranks in general-language corpora.

This applied primarily to high-frequency lexical items (e.g., *я* ‘I’, *працювати* ‘to work’, *добрий* ‘good’) and function words (*у/в* ‘in’, *але* ‘but’, *бо* ‘because’), where all three indicators aligned.

Discrepancies between textbook and general corpus frequencies typically related to food names (*молоко* ‘milk’, *картопля* ‘potato’, *борщ* ‘borscht’), city locations (*готель* ‘hotel’, *кафе* ‘cafe’), culturally relevant items or curriculum-based vocabulary (detailed below). These were frequent in textbooks but rare in general corpora, justifying lower-level assignments due to communicative priority in UFL contexts.

Classifying curriculum-based vocabulary posed a particular challenge, as it included two categories: curriculum-related and curriculum-defined vocabulary. The former consists of terms like *зошит* ‘workbook’, *правило* ‘rule’, or *дієслово* ‘verb’, which are frequent in educational settings but not in general-language corpora. The latter balances word frequency with cognitive load in learning a synthetic language. Experts occasionally grouped words by grammatical functions introduced at a given level, for example, verbs like *захоплюватися* ‘be fond of’, *цікавитися* ‘be interested in’, and *займатися* ‘practice’ were assigned to B1 to reflect their role in teaching the instrumental case.

Further differences emerged between corpus data and expert judgment. Despite high general-language frequency, abstract nouns (*наука* ‘science’, *освіта* ‘education’, *мета* ‘purpose’) and emotion terms were placed at B1 level due to conceptual complexity, aligning with CEFR recommendations. Expert disagreement also arose over internationalisms: while one expert favored A1 placement for words like *стейк* 'steak' or *емоційний* 'emotional', the other advocated for gradual distribution across A1–B1.

For lower-frequence items (below position 1,700), with ≤14 occurrences in A1–A2 textbooks, general corpus data became more relevant. These words were assigned to A2 or B1 levels based on three guiding principles:

1) frequency ranks in general-language corpora,

2) the significant onset of use, accounting for frequency and dispersion across textbook levels, and

3) the alignment of didactic requirements with CEFR descriptors.

Items beyond position 2,500 were predominantly assigned to B1, though exceptions remained (e.g., *петрушка* 'parsley', *кекс* 'cupcake' at A2) when their communicative relevance justified earlier introduction.

Overall, the thematic, semantic, and derivational grouping of items, applied during the internal alignment stage, further supported consistent level assignments and helped reconcile expert differences. However, during the internal level of alignment, educational factors, including the learning environment and objectives that define the learner's priorities and competences, played a role in labeling curriculum-based vocabulary. Additionally, the external alignment with CEFR and UFL Standard encountered several challenges: linguistic, stemming from the complexities of Ukrainian as a synthetic language, and context-related, primarily sociolinguistic, associated with the usage of specific words in particular contexts, e.g. feminine counterparts.

Therefore, discrepancies in expert classification—which relied primarily on frequency and dispersion data from L2 textbooks, frequency ranks in general corpora, and examples of usage from both textbooks and corpora—were rooted in the experts' consideration of the cognitive load on the learner, the grammatical complexity of a lexical unit, and the instructional requirements for a specific competency level.

4.3 Vocabulary Coverage and Level Distribution

Our analysis revealed key challenges in Ukrainian language resources, including inconsistencies between textbooks and corpora, high lexical variability, and a predominance of low-frequency vocabulary (80% of 37,087 unique lemmas occur only 1–10 times in a one-million-word corpus). The overrepresentation of low-frequency vocabulary in UFL textbooks reduces opportunities for repeated encounters with core vocabulary, potentially hindering lexical retention, reading fluency, and overall language acquisition. Additionally, inconsistent CEFR alignment results in level-straddling textbooks that blur level boundaries. To address these issues, we developed a CEFR-aligned vocabulary list for UFL using a multifaceted approach that combines corpus frequency and dispersion data, expert review, and semi-automatic thematic, semantic, and derivational grouping, ensuring a comprehensive and systematic vocabulary profiling.

At this stage, 5,891 lemmas from A1–B1 textbooks have been CEFR-labeled. As Table 10 shows, 969 words were assigned to A1 (vs. 700 expected), 1,394 to A2 (vs. 1,000 expected), and 2,141 to B1 (vs. 1,500 expected). From this list, 1,100 word families were formed. In the future, this process will be extended to B1–B2 and C1–C2 data to further refine the distribution of words across levels.

The selected 2,363 lexical items at A1–A2 levels cover 48% of running words in fiction texts, 40% in news, and 32% in academic texts in the GRAC-17a corpus (2000–2023). In contrast, the English General Service List (1953) provides approximately 82% average coverage across genres, including 90.6% for short novels and 78.4% for academic texts (Nation & Waring, 1997). This gap may stem from lower lexical control in UFL textbooks, typological differences, and the high frequency of compounds in English. These findings highlight the need for further research on vocabulary coverage in Ukrainian.

Level	Expected		Received	
	per level	Total	per level	Total
A1	700	700	969	969
A2	1000	1700	1394	2363
B1	1500	3200	2141	4504
B2	1900	5100	1140	
C1	2300	7400	220	
C2	2600	10000	27	

Table 10: Vocabulary Distribution in the Ukrainian Profile Across A1-B1 Levels

The selected words are categorized into 39 thematic groups. At A1, the largest groups are Food and drinks (91 units), Training and education (75), Communication (66), Time (65), Numbers (64), and Actions (48). Figure 2 compares A1 and A2 levels, revealing vocabulary distribution within themes. The most notable increase occurs in Food and drinks (91 to 137 items), Description of things (27 to 84), Actions (48 to 84), Personal feelings (7 to 36), and Body and health (25 to 55).

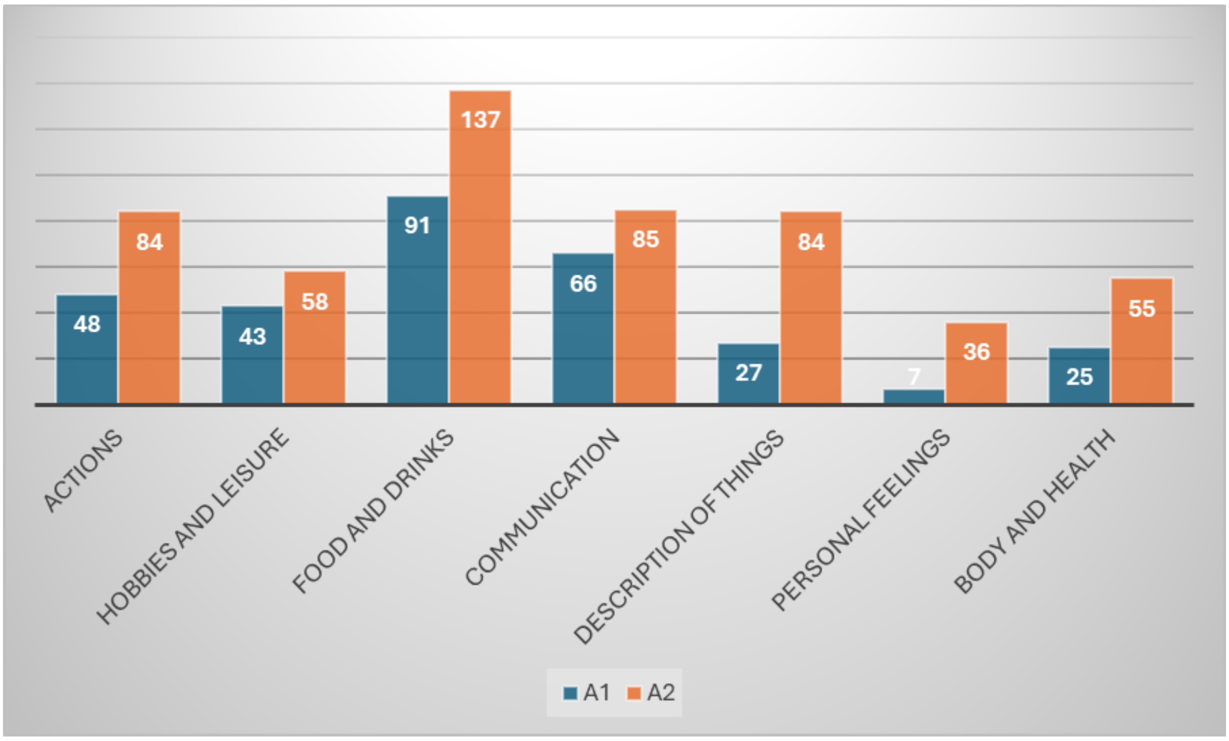


Figure 2: Vocabulary Distribution within Thematic Groups Across A1-A2 Levels

Figure 3 below illustrates the distribution of parts of speech across A1-A2 levels, showing an increase at A2 in verbs (120 to 241 items), adjectives (86 to 258), nouns (534 to 581), and adverbs (81 to 118), as well as a decrease in the number of numerals (61 to 20), pronouns (26 to 12), and function words (46 to 33).

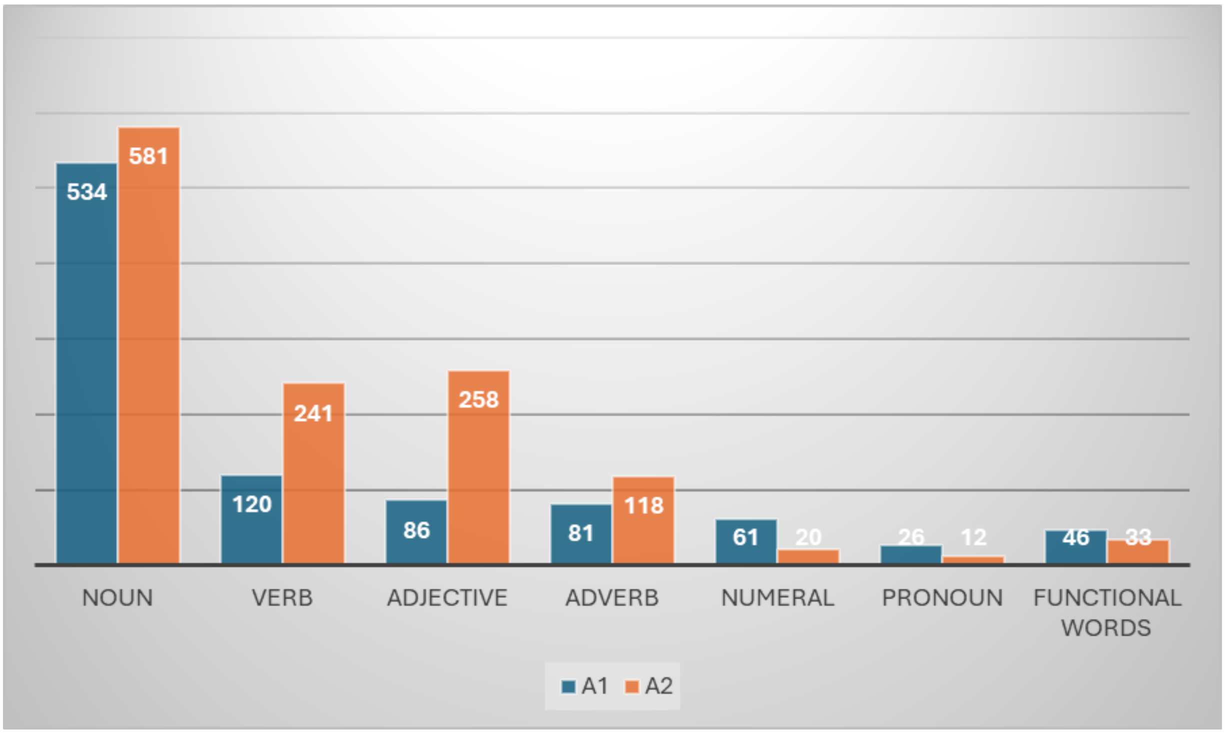


Figure 3: Part-of-Speech Distribution Across A1 and A2 Levels

4.4 PULS Platform

The selected and CEFR-labeled items are published on the PULS platform (PULS, 2025), a digital learning resource with integrated lexical database functionality. Users can currently filter the vocabulary by level, part of speech, and thematic group, with the guide word and word family information to be added in future updates. At present, vocabulary items for levels A1–B1 are available, with content for higher levels under development. This profile serves as the foundation for the prospective Ukrainian Learner's Dictionary (ULD), which will include detailed lexical entries with part of speech, CEFR label, thematic group, definition at the level of individual senses, corpus-based examples, pronunciation (audio), English equivalents, pictorial illustrations where relevant, and semantic and derivational relations.

4.4.1 Software

A Ruby on Rails web application was developed to automate vocabulary list management, streamlining the processing of thousands of lexical items while incorporating teacher feedback for continuous refinement. The platform uses PostgreSQL for database management and a front-end built with HTML, CSS, and JavaScript.

5. Conclusions and Future Plans

The proposed multifaceted approach that combines corpus frequency and dispersion data, expert review, and lexical analysis within word groups, has proven effective in developing a data-based CEFR-aligned vocabulary list for UFL. To date, 5,891 lemmas have been selected, with a target of 10,000 lemmas to be distributed across CEFR levels. Based on semantic and derivational relations, the selected vocabulary is organized into 39 thematic groups and 1,100 word families. The results are published and visualized on the PULS platform.

In the future, we plan to further refine CEFR-level distribution of the selected lexical items during standard-setting procedures. To this end, we plan to build a balanced reference corpus of Ukrainian and a learner corpus based on student proficiency exam essays. Together with the CEFR-aligned word list, these resources will serve as the foundation for a future UFL Learner's Dictionary. This will be the first-ever CEFR-labeled corpus-based UFL reference source that will serve the needs of learners, educators, material creators, and proficiency test designers.

6. Acknowledgements

This project has been supported by a grant from UCU's research fund. The ideas for this project were developed in consultation with Jelena Kallas from The Institute of Estonian Language and Matej Klemen and Špela Arhar Holdt from University of Ljubljana. We would also like to extend our heartfelt thanks to Nazar Danylkiv, Director of the School of Ukrainian Language and Culture at UCU, as well as to the

School's managers Natalia Kravets and Khrystyna Popovych, for their encouragement and assistance. We are also grateful to Oleksandra Safatiuk and Sofia Oseredchuk, students at UCU's Philology Program, for their help in developing materials for the PULS platform.

7. References

Alfter, D. (2021). Exploring Natural Language Processing for Single-Word and Multi-Word Lexical Complexity from a Second Language Learner Perspective. *Data Linguistica*, 31, ed. by Lars Borin.

Alfter, D., Bizzoni, Y., Agebjórn, A., Volodina, E., and Pilán, I. (2016). From Distributions to Labels: A Lexical Proficiency Analysis Using Learner Corpora. In *Proceedings of the Joint Workshop on NLP4CALL and NLP for Language Acquisition at SLTC*, 130, pp. 1–7.

Aligning Language Education with the CEFR: A Handbook. (2022). British Council, ALTE.

B1 Preliminary Vocabulary List. (2020). Cambridge Assessment English. Accessed at: https://www.cambridgeenglish.org/images/506887-b1-preliminary-2020-vocabulary-list.pdf. (1 June 2025)

Bartkiv, N. & Borodin, K. (2022). *Yabluko: pidruchnyk z ukrayinskoyi movy yak inozemnoyi (rubizhnyy riven).* Vydavnytstvo UKU, Lviv.

Borodin, K. & Turkevych, O. (2023). The "1000 and 1 Words" (A1). *Slavic Language Education*, 3 (Language-Lab). Accessed at: https://doi.org/10.18452/28236. (1 June 2025)

Breakthrough*: An Objective at Level A1 of the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR).* (2009). Strasbourg: Council of Europe. Accessed at: https://ealta.eu/documents/resources/Breakthrough%20specification.pdf. (1 June 2025)

BRUK: *Ukrainian Brown Corpus.* Accessed at: https://github.com/brown-uk/dict_uk. (7 July 2025)

Buk, S. (2006a). *3 000 naychastotnishykh sliv naukovoho styliu suchasnoyi ukrayinskoyi movy.* Nauk. red. F. Batsevych. Lviv: LNU imeni Ivana Franka.

Buk, S. (2006b). *3 000 naychastotnishykh sliv rozmovno-pobutovoho styliu suchasnoyi ukrainskoyi movy.* Nauk. red. F. Batsevych. Lviv: LNU imeni Ivana Franka.

Buk, S. (2006c). Chastotnyi slovnyk ofitsiyno-dilovoho styliu: pryntsypy ukladannia ta statystychni kharakterystyky. In *Linhvistychni studiyi*: Zb. nauk. prats, 14, pp. 184–188.

Burak, M. (2015). *Yabluko: pidruchnyk z ukrayinskoyi movy yak inozemnoyi (bazovyy riven).* Vydavnytstvo UKU, Lviv.

Capel, A. (2010). A1-B2 Vocabulary: Insights and Issues Arising from the English Profile Wordlists Project. *English Profile Journal*, 1(1), pp. 1–11.

Common European Framework of Reference for Languages: *Learning, Teaching, Assessment. Companion volume.* (2020). Strasbourg: Council of Europe Publishing.

Coxhead, A. (2011). The Academic Word List 10 Years on: Research and Teaching Implications. *Tesol Quarterly* 45 (2), pp. 355–362.

EVP: *English Vocabulary Profile.* Cambridge University Press & English Profile. Accessed at: https://www.englishprofile.org/wordlists. (7 July 2025)

Green, A. (2010). Requirements for Reference Level Descriptions for English. (2010). *English Profile Journal*, 1(1), pp. 1–19.

Kaczmarek, H. (2006). Minimum leksykalne a orientacja komunikacyjna w nauczaniu języka obcego. Kilka uwag na przykładzie niemieckiego jako języka obcego. In *Studia Neofilologiczne*, z. V, s. 27–35.

Kardela, H. (2015). Lexical Nests Revisited: a Cognitive Grammar Account. *SKASE Journal of Theoretical Linguistics.* Vol. 29, pp. 292–312.

Klemen, M., Arhar Holdt, Š., Pollak, S., Kosem, I., Pori, E., Gantar, P., Knez, M. (2023). Building a CEFR-Labeled Core Vocabulary and Developing a Lexical Resource for Slovenian as a Second and Foreign Language. In *Electronic lexicography in the 21st century.* Brno, pp. 664-678.

Nation, P. (2000). Learning Vocabulary in Lexical Sets: Dangers and Guidelines. *TESOL Journal*, pp. 6–10.

Nation, P. & Waring, R. (1997). Vocabulary Size, Text Coverage and Word Lists. In N. Schmitt & M. McCarthy (eds.) *Vocabulary: Description, Acquisition, and Pedagogy.* Cambridge: Cambridge University Press, pp. 6–19.

NLP-UK: *NLP-UK toolkit for Ukrainian.* Accessed at: https://github.com/brown-uk/nlp_uk. (7 July 2025)

O'Sullivan, B. (2021). *The Comprehensive Learning System.* British Council.

Partyko, Z. (2004). *Slovnyk-minimum ukrayinskoyi movy.* Kyiv.

Pastuchowa, M. (2009). O słowotwórstwie z perspektywy leksykalnej. In Achtelik, Aleksandra & Tambor, Jolanta (eds.). *Sztuka czy rzemiosło? Nauczyć Polski i polskiego.* Katowice: Wydawnictwo Gnome, pp. 21–27.

Pintard, A. & François, T. (2020). Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words. In *Language Resources and Evaluation Conference*, pp. 85–92.

PULS platform: *Ukrainian Vocabulary Profile.* (2025). Lviv. Accessed at: https://puls.peremova.org/. (7 July 2025)

Shvedova, M., Waldenfels, R. von, Yarygin, S., Rysin, A., Starko, V., Nikolajenko, T. et al. (2017–2024). *GRAC: General Regionally Annotated Corpus of Ukrainian.* Electronic resource: Kyiv, Lviv, Jena. Accessed at: http://uacorpus.org. (7 July 2025)

Shvedova, M. (2020). The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality. In *Computational*

Linguistics and Intelligent Systems. Proc. 4th Intl. Conf. COLINS 2020, pp. 489–506.

Starko, V. & Rysin, A. (2022). VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool. In *Computational Linguistics and Intelligent Systems. Proc. 6th Int. Conf. COLINS 2022*, Gliwice, Poland, pp. 71–80.

Starko, V. & Rysin, A. (2023). Creating a POS Gold Standard Corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop* (UNLP). ACL, pp. 91–95.

Synchak, O. (2015). *Yabluko: pidruchnyk z ukrayinskoyi movy yak inozemnoyi (vyshchyy riven)*. Vydavnytstvo UKU, Lviv.

SLSUFL: *The State Language Standard "Ukrainian as a Foreign Language. Levels of General Proficiency A1-C2"* (2024). Accessed at: https://unity.gov.ua/en/2024/09/06. (7 July 2025)

UFLTC: *The UFL Textbook Corpus*. Accessed at: http://corpus-puls.peremova.org. (7 July 2025)

Üksik, T., Kallas, J., Koppel, K., Tsepelina, K., Pool R. (2021). Estonian as a Second Language Teacher's Tools. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 130–134.

VESUM: Rysin, A. & Starko, V. *Large Electronic Dictionary of Ukrainian (VESUM)*. (2005-2025) Web version 6.6.10-SNAPSHOT. Accessed at: https://vesum.nlp.net.ua. (7 July 2025)

Volodina, E., Alfter, D., Lindström Tiedemann, T. (2024). Profiles for Swedish as a Second Language: Lexis, Grammar, Morphology. In *Proceedings of the Huminfra conference*, pp. 10–19.