Ukrainian Multiword Expressions Corpus: Creation, Annotation, and Linguistic Analysis

Hanna Sytar
Institute of Slavonic Studies of the Czech Academy of Sciences (ISS CAS), Valentinska 91/1, 11000 Praha
hanna.sytar@slu.cas.cz

Maria Shvedova
National Technical University "Kharkiv Polytechnic Institute"
Kyrpychova str. 2, 61002, Kharkiv
mariia.shvedova@khpi.edu.ua

Olha Kanishcheva
Heidelberg University
Grabengasse 1, 69117 Heidelberg
SET University
Mykoly Shpaka St. 3, 03113, Kyiv
kanichshevaolga@gmail.com

Abstract

This paper presents the development of a corpus of annotated multiword expressions (MWEs) for Ukrainian. The resource covers four major categories of MWEs: verbal, nominal, adjectival/adverbial, and functional. We describe the methodology used for data selection, the annotation scheme, and the procedures employed during annotation. In addition, the paper discusses some specific types of MWE constructions, illustrating their usage with numerous examples and addressing complex and borderline cases. The resulting corpus is an important resource for linguistic studies and NLP tasks involving MWEs, and is publicly accessible here.

1 Introduction

Multiword expressions (MWEs) – such as idioms, light verb constructions, and collocations – play a crucial role in both linguistic theory and natural language processing (NLP) (Constant et al., 2017; Giouli and Barbu Mititelu, 2024). They represent combinations of words whose meaning cannot always be reasoned from their components (non-compositional), and they are essential for accurate parsing, translation, and lexical semantics. The identification and correct processing of MWEs have been shown to improve performance across a wide range of NLP tasks, including machine translation, information extraction, and language modeling. Therefore, the availability of high-quality MWE-annotated corpora is very important for the development of language technologies that can handle idiomatic and non-compositional constructions effectively (Savary et al., 2017).

For the Ukrainian language, MWE research continues to be fairly underexplored. However, despite recent progress in Ukrainian NLP, there is still a lack of systematically annotated data that represent the diversity and complexity of MWEs. This gap is partly due to the linguistic characteristics of Ukrainian – a morphologically rich and syntactically flexible language, where free word order, inflectional variation, and the presence of MWE variants make automatic identification of MWEs particularly challenging. Moreover, the lack of existing linguistic resources limits the ability to train and evaluate computational models for Ukrainian MWE detection.

In this article, we describe our experience in creating, annotating, and analyzing a new corpus of Ukrainian multi-word expressions as part of the multilingual PARSEME shared task[1]. The corpus includes manually annotated MWEs and is designed to support both linguistic research and computational modeling.

This article is structured as follows. Section 2 discusses how phenomena corresponding to multiword expressions in the PARSEME framework are treated within different areas of traditional Ukrainian linguistics, including phraseology, and different branches of grammar. Section 3 describes the corpus structure and data sources. Section 4 outlines the annotation scheme and process, as well as the types of MWEs, with examples. Section 5 addresses complex and borderline cases, including specific constructions not fully covered by the PARSEME scheme, such as multiword particles, challenging instances of inherently adpositional verbs, multiword adpositions, and the variation observed within MWEs, based on the results of the corpus analysis. Section 6 concludes the article and outlines plans.

2 Multiword Expressions in Ukrainian Linguistics

In Ukrainian linguistics, the term MWE is still not widely used. Different types of MWEs are studied within both phraseology and grammar.

Firstly, Ukrainian phraseology traditionally ap-

[1] https://gitlab.com/parseme/corpora/-/wikis/home

plies a narrow approach to fixed expressions, which includes only phraseologisms proper (idioms). Proverbs and sayings, language clichés, etiquette formulas, phraseme-like constructions, and other fixed expressions are not included in this analysis (Alefirenko, 1988; Bilonozhenko et al., 1993; Uzhchenko and Uzhchenko, 2005). Accordingly, well-known Ukrainian phraseological dictionaries contain only idioms (Bilonozhenko et al., 2003, 1993), while proverbs and sayings are compiled separately (Nomys, 1993). Under this traditional approach, most types of MWEs are not represented in lexicographic resources and are not considered in the development of computational tools.

Secondly, various types of MWEs are partially described across separate branches of Ukrainian grammar under different terminological labels. For example, light verb constructions (LVC) correspond to *periphrastic verb-noun constructions* or *periphrastic predicates*, studied within functional-communicative or semantic syntax (Zahnitko, 2001; Sytar, 2010). The term adposition idiom (AdpID) corresponds to what is called a *secondary compound preposition* in Ukrainian grammar (Vykhovanets', 1980; Vykhovanets' et al., 2017), or alternatively *prepositional equivalent* or *prepositional analogue* (Luchyk, 2006; Zahnitko et al., 2007; Kushch, 2008; Zahnitko et al., 2009). Conjunction idioms (ConjID), known as *secondary compound conjunctions*, are recognized as a distinct structural type of conjunctions (Vykhovanets' et al., 2017) and are described lexicographically in (Horodens'ka, 2007; Luchyk, 2006).

The definition of inherently adpositional verbs (IAV) is closely related to the well-developed concept of verbal valency in Ukrainian grammar, including valency-determined obligatory argument positions of the verbal predicate (Vykhovanets', 1988; Zahnitko, 1996; Masyts'ka, 1998) and the concept of verbal government, which has also been lexicographically documented (Kolibaba and Fursa, 2025).

3 Corpus Design and Data Sources

Research on multiword expressions has attracted increasing attention in recent decades, with multilingual NLP initiatives – such as the PARSEME shared tasks (Savary et al., 2017; Ramisch et al., 2020) – establishing common typologies and annotation standards for over 30 languages. These efforts have produced multilingual corpora that now serve as essential benchmarks for automatic MWE processing.

For Slavic languages, existing resources (e.g., for Polish, Czech, and Bulgarian) demonstrate that rich morphology and flexible syntax consistently complicate both annotation and automatic detection (Savary and Waszczuk, 2020; Stoyanova et al., 2016; Pala et al., 2008). Until now, Ukrainian has lacked a fully comprehensive systematically annotated MWE corpus. UD_Ukrainian-ParlaMint (Shvedova et al., 2025) contains MWE information partially through *fixed* dependency relations, with heads annotated using *ExtPos* tags (external POS feature indicating the effective part of speech of an expression); however, this annotation covers only two PARSEME categories (adjectival/adverbial MWEs and functional MWEs)[2].

3.1 Data Sources

All annotated data originate from the General Regionally Annotated Corpus of Ukrainian (GRAC)[3] (Shvedova, 2020). The selected texts come from the *Ukrainian Week* newspaper (2013-2016), and the data type is interview. Initially, the corpus was automatically annotated using the UDPipe 2 model for Ukrainian (ukrainian-iu-ud-2.15-241121)[4](Straka, 2018), providing lemmas, UPOS and XPOS tags, and morphological features in accordance with Universal Dependencies conventions. Multiword expressions were then manually annotated with the FoLiA Linguistic Annotation Tool (FLAT)[5], following the PARSEME MWE 2.0 guidelines.

3.2 Corpus Statistics

The current version of the corpus contains 12,078 sentences with a total of 198,555 tokens, including 5,993 annotated multiword expressions. Each document is enriched with metadata detailing its source, genre, and publication year. Annotations are provided in CoNLL-U format, ensuring compatibility with Universal Dependencies resources and other corpus analysis tools.

The annotated MWEs are categorized as follows: verbal – 2,804, nominal – 818, adjectival and adverbial – 1,017, functional – 1,354, deverbal nouns – 345, and idioms – 1,134.

[2]https://universaldependencies.org/uk/feat/ExtPos.html
[3]https://uacorpus.org/
[4]https://ufal.mff.cuni.cz/udpipe/2/models
[5]https://flat.readthedocs.io/

It is important that our data consists of contemporary journalistic texts containing newly emerging multiword expressions that have not been documented in phraseological dictionaries. Ukrainian phraseological dictionaries were compiled in the late 20th and early 21st centuries, with their primary sources being folklore and works of Ukrainian literature from the 19th to 20th centuries. Examples of such new expressions include *vnutrišn'o peremiščena osoba* 'internally displaced person', *tymčasovo okupovana terytorija* 'temporarily occupied territory', *zeleni čolovičky* 'little green men' (unmarked soldiers), *hlyboka sturbovanist'* 'deep concern' (diplomatic formality masking inaction).

The corpus also contains colloquial variants of fixed expressions, e.g., *vymušenyj pereselenec'* 'forced migrant', *povna majačnja* 'complete nonsense', *vse po fen-šuju* 'everything as it should be' (lit. 'everything according to Feng Shui'), *vidpravyty na try litery* 'tell someone to go to hell' (lit. 'send to three letters').

At the same time, the annotated MWE set includes prepositional units that have not previously been described in Ukrainian grammars or dictionaries, such as *komitet* ***u spravax*** *nacional'nostej* 'committee **on** nationalities'; ***na moment*** *svoho vidkryttja* '**at the time of** its opening'; *Riven' ïï pidtrymky kolyvajet'sja* ***v korydori*** *60-70%* 'its support level fluctuates **between** 60 and 70%'.

Additionally, cases have been observed where the meaning of well-known idioms has shifted; e.g., *imperija zla* 'evil empire', a phrase used by Ronald Reagan in a 1983 speech to refer to the USSR, is used in Ukrainian texts of recent years to denote Russia as a country that continues the totalitarian and imperial policies of the Soviet Union: *– Why do so few Russians sympathize with the Maidan? – Because Russia is an empire.* ***An evil empire.*** *A fragment of the Soviet Union, not yet ready for something different. They want to rule over others; the empire is still coursing through their blood. (Ukrainian Week, 2014; our transl. from Ukr.)*

Therefore, the created corpus can partly compensate for the incompleteness of existing phraseological and grammatical dictionaries of Ukrainian and serve as a valuable resource for addressing various NLP tasks.

4 Annotation Scheme

The annotation scheme for the Ukrainian MWE corpus follows the general principles of the PARSEME Shared Task 2.0 guidelines[6], with adaptations that reflect the specific grammatical and lexical properties of Ukrainian. The goal of the scheme is to maintain cross-linguistic compatibility while accurately capturing constructions characteristic of Ukrainian. Figure 1 presents an example of an output file from our corpus and illustrates the corresponding format.

The Ukrainian MWE corpus is distributed in the standard .cupt format. The linguistic annotation follows the cupt column structure. Lemmas (column 3), UPOS tags (column 4), XPOS tags (column 5), morphological features (column 6), as well as syntactic heads and dependency relations (columns 7-8), are automatically generated using UDPipe 2. The UPOS and FEATS columns follow the Universal Dependencies tagsets, while XPOS is likely based on the AnCora tagset. Additional metadata in the MISC column (column 10) is also automatically provided. The PARSEME:MWE column (column 11) contains manually assigned labels for multiword expression categories, including VID, LVC.full, LVC.cause, IRV, and the experimentally annotated IAV category. All automatic annotations were produced using the UDPipe 2 model[7].

In the following, we describe the main MWE categories and subcategories, together with representative examples of multiword expressions in Ukrainian.

4.1 MWE Types

The top-level categories cover all syntactic types of MWEs and include verbal MWEs (VMWEs), nominal MWEs (NMWEs), adjectival and adverbial MWEs (AMWEs), and functional MWEs (FuncMWEs). This comprehensive classification is introduced in version 2.0 of the annotation guidelines, extending earlier versions of PARSEME that covered verbal MWEs only.

During manual annotation, candidate multiword expressions are classified using category-specific decision diagrams. The annotation scheme distinguishes four major MWE classes: verbal, nominal, adjectival/adverbial, and functional.

Verbal MWEs (VMWEs) are subdivided into universal, quasi-universal, language-specific, and

[6]https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/
[7]https://ufal.mff.cuni.cz/udpipe/2/models

```
# source_sent_id = . . news-69-27
# text = У той самий час наша демократія розпадається на друзки.
1   У    у    ADP Spsa    Case=Acc    4   case    _   _   *
2   той той DET Pd--mnsaa   Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing|PronType=Dem  4   det _   _   1:DetID
3   самий   самий   DET Px--mnsaa   Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing|PronType=Prs|Reflex=Yes   4   det _   _   1
4   час час NOUN    Ncmsan  Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing   7   obl _   _   *
5   наша    наш DET Ppslf-sna   Case=Nom|Gender=Fem|Number=Sing|Person=1|Poss=Yes|PronType=Prs  6   det _   _   *
6   демократія  демократія  NOUN    Ncfsnn  Animacy=Inan|Case=Nom|Gender=Fem|Number=Sing    7   nsubj   _   _   *
7   розпадається    розпадатися VERB    Vmpip3s Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin    0   root
8   на  на  ADP Spsa    Case=Acc    9   case    _   _   2
9   друзки  друзка  NOUN    Ncmpan  Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur   7   obl _   SpaceAfter=No   2
10  .   .   PUNCT   U   _   7   punct   _   _   *
```

Figure 1: The output format of the Ukrainian MWE corpus is illustrated below using a sample sentence: *U toj samyj čas naša demokratija rozpadajet'sja na druzky.* 'At the same time, our democracy is falling apart.' In this example, the multi-word expressions *toj samyj* 'the same' (DetID) and *rozpadajet'sja na druzky* 'falling apart' (VID) are annotated (MWEs are in red blocks).

an optional experimental category. Universal VMWEs include light verb constructions and verbal idioms, while quasi-universal categories cover inherently reflexive verbs, idiomatic verb-particle constructions, and multi-verb constructions. Language-specific categories are defined separately for each language, and an experimental category is introduced for inherently adpositional verbs.

Nominal MWEs (NMWEs) comprise nominal idioms, pronominal idioms, and deverbal nominal MWEs derived from verbal MWEs, inheriting their subcategorization.

Adjectival and adverbial MWEs (AMWEs) include adjectival idioms, adverbial idioms, and deverbal MWEs derived from verbal constructions.

Finally, **functional MWEs (FuncMWEs)** form a universal class and include determiner, adposition, conjunction, and interjection idioms.

More details about the MWE subtypes and examples can be seen in Table 1.

This typology ensures comprehensive coverage of syntactic and functional MWE types, providing a consistent framework for manual annotation and supporting subsequent computational processing.

Fig. 2 shows the distribution of Ukrainian multi-word expression types by frequency. The vertical axis lists the MWE types, while the horizontal axis represents the number of occurrences of each type in the corpus.

The distribution is uneven and highly skewed. The most frequent type is IAV, which clearly dominates all other categories with more than 1,200 instances. Other high-frequency types include LVC.full, AdpID, and AdvID, each represented by several hundred occurrences.

A noticeable but lower frequency is observed for VID and NID, which form a medium-frequency group. In contrast, many MWE types (such as AV.LVC.cause, IVPC.full, MVC, and NV.VID) are represented by only a few instances.

Overall, the diagram demonstrates a long-tail distribution typical of linguistic data: a small number of MWE types account for the majority of occurrences, while most types occur rarely.

4.2 Annotation Process

Annotation was performed manually by two linguists using the FLAT annotation platform[8]. The annotators followed detailed written guidelines derived from the PARSEME framework[9]. Ambiguous cases and borderline expressions were discussed collaboratively to ensure consistency.

4.3 Quality Control and Agreement

To assess annotation reliability, a subset of 20 files was independently annotated by two researchers. Inter-annotator agreement (IAA) was calculated using the MWE-based F-measure (Savary et al., 2017), resulting in an MWE-based F-score of 54. Disagreements were resolved through discussion and guideline analysis.

The evaluation of MWE annotation shows significant variation across categories, with functional expressions showing the highest reliability. AdpID and AdvID achieved the most robust results, with F1-scores of 0.845 and 0.754 respectively, suggesting that adpositional and adverbial idioms are more easily identifiable in Ukrainian. Conversely, verbal constructions such as IAV and LVC.cause suffer from a severe *recall gap* (0.158 and 0.156), where high precision indicates that while annotations are accurate, a vast majority of instances remain undetected. More detailed information about the evaluation scores for each MWE class can be seen in Figure 3.

[8] https://flat.readthedocs.io/en

[9] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

Type / Subtype	Examples / Description
LVC.full	Semantically bleached verb (e.g., *vyslovyty zaperečennja* 'to raise an objection')
LVC.cause	Verb adds causative meaning (e.g., *spryčynyty rujnuvannja* 'to cause destruction')
VID	Verbal idioms (e.g., *skakaty v hrečku* 'to commit adultery', lit. 'to jump into buckwheat')
IRV	Inherently reflexive verbs (e.g., *dozvolyty sobi* 'to afford', lit. 'to allow oneself')
IVPC.full / IVPC.semi MVC	Multi-verb constructions (e.g., *daty (komus') zrozumity* 'let someone know / make clear')
IAV (experimental)	Inherently adpositional verbs (e.g., *vplyvaty na* 'influence smth.')
NID	Nominal idioms (e.g., *prymxa doli* 'whim of fate')
PronID	Pronominal idioms (e.g., *odyn odnoho* 'one another')
NV	Deverbal nominal MWEs derived from VMWEs (e.g., *znjattja sankcij* 'lifting of sanctions')
AdjID	Adjectival idioms (e.g., *tak zvanyj* 'so-called')
AdvID	Adverbial idioms (e.g., *ostannim časom* 'recently', lit. 'in recent times')
AV	Deverbal AMWEs derived from VMWEs (e.g., *ozbrojenyj do zubiv* 'heavily armed', lit. 'armed to the teeth')
DetID	Determiner idioms (e.g., *toj čy inšyj* 'a particular', lit. 'one or another')
AdpID	Adposition idioms (e.g., *pid čas* 'during', lit. 'under time')
ConjID	Conjunction idioms (e.g., *dlja toho, ščob* 'in order to' lit. 'for that to')
IntjID	Interjection idioms (e.g., *Slava Bohu!* 'Thank God!')

Table 1: Classification of multiword expression types with their main categories and examples.

5 Discussion

In this section, we discuss several notable features and challenges encountered during the annotation of Ukrainian MWEs, highlighting patterns that may be relevant for other Slavic languages and suggesting potential extensions to the existing classification framework.

5.1 Particle Idioms

Our annotation experience with Ukrainian MWEs suggests that the current classification employed in the project would benefit from the inclusion of an additional type, **Particle Idioms (PartID)**. The news corpus contains a considerable number of multiword particles that, during annotation, were assigned to the *Other* category: *vse ž taky* 'after all / still / nevertheless', *navrjad čy* 'hardly / unlikely', *xiba ščo* 'unless / except perhaps', *xoč by* 'at least', *xoča b* 'at least', etc.

We assume that multicomponent (compound) particles are not specific to Ukrainian alone (Zahnitko and Karataieva, 2012), but are also characteristic of other Slavic languages; cf. Czech *kéž by* 'if only / I wish', *ještě aby* 'as if (… were to)': Czech.: ***Kéž by** se mu to povedlo!* 'May he succeed in this!' ***Ještě aby** si stěžoval!* Lit. 'As if he were to complain!'; idiomatic meaning: 'He has no right to complain.' Polish: *Trzeba próbować, **a nuż** się uda?* 'You should try, **what if** it works?' *Płowa zwierzyna to **bądź co bądź** zwierzyna szlachetna.* 'An ungulate is, **after all**, a noble animal', lit. *bądź co bądź* 'be what be'.

During annotation, cases were found where combinations such as *particle+preposition*, *pronoun+preposition*, etc. are used as compound particles, i.e., in contemporary Ukrainian, they function as **Particle Idiom (PartID)**: *Jakščo ljudyna xoče provezty vodu, ščodennyky, olivci, to **do čoho** tut Služba bezpeky?* 'If a person wants to transport water, diaries, pencils, **what** does the Security Service **have to do with** it?' ***Ščo za** dyvyna taka xovajet'sja za cym terminom, my šče pohovorymo nyžče.* '**What kind of** wonder lies behind this term, we will discuss below.' *Novyny ne dyvljusja – ščos' meni **ne do** nyx.* 'I don't watch the news – I am **not in the mood for** it'.

These cases are a zone of intersection between MWEs and phraseme constructions. In the COST Action CA22115[10] Memorandum is indicated that phraseme construction (PhraCons) is a construction that "consist of one or more lexically fixed

[10]https://www.phraconrep.com/

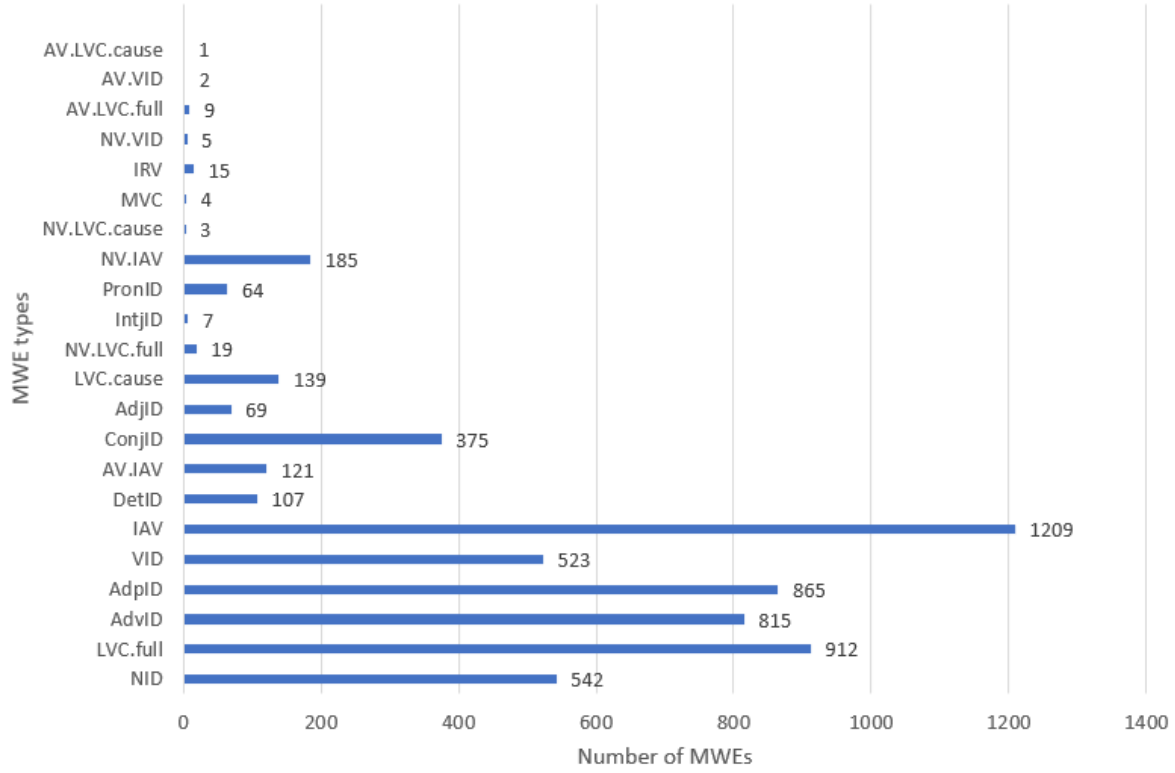

Figure 2: Distribution of Ukrainian MWE types by frequency.

element/s (=anchor/s) and one or more open slot/s (Dobrovol'skij, 2011). The slots must be filled by lexical elements (=fillers) according to lexical, grammatical, communicative, stylistic and intonational rules. Although PhraCons are partially schematic, they have an abstract overall meaning that is usually idiomatic, which means it cannot be attained simply by adding up the meanings of its constituents". This specific type of construction is the focus of *COST Action CA22115* - A Multilingual Repository of Phraseme Constructions in Central and Eastern European Languages (PhraConRep) (Braxatorisová, 2024), where Ukrainian is among the 15 languages under study. The structural, semantic, and pragmatic properties of phraseme constructions in Ukrainian from the perspective of construction grammar are described in (Sytar, 2017). It should be noted that in such contexts the phraseme construction is broader than the MWE: the phraseme construction corresponds to the pattern N_{dat} *ne do* N_{gen} (*Meni* ***ne do*** *novyn.* 'I am **not in the mood for** news.'), whereas the MWE is limited to *ne do*.

5.2 Inherently Adpositional Verb (IAV)

As shown in Figure 2, our text corpus revealed 1,209 contexts of inherently adpositional verbs (IAVs), which constitutes the absolute majority, exceeding the predictably frequent nominal, adverbial, and verbal idioms (912, 815, and 523, respectively). Cases of special optional and experimental inherently adpositional verbs (or prepositional verbs) caused the greatest difficulties and required discussion and agreement among annotators.

According to the project documentation, this MWE type encompasses two groups of cases: "It consists of a verb or VMWE and an idiomatic selected preposition or postposition that is either always required or, if absent, changes the meaning of the verb or VMWE significantly."[11] Both subtypes are present in Ukrainian:

a) verbs with mandatory postverbal prepositional complement: *asocijuvatysja z*+Ins 'to be associated with smth.', *vplyvaty na* +Acc 'to influence smth.', *gruntuvatysja na*+Dat 'to be based on smth.', *narażatysja na*+Acc 'to face smth.'.

b) polysemous verbs, which have different meanings with and without a prepositional complement:

[11] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/index.php?page=iav#iav

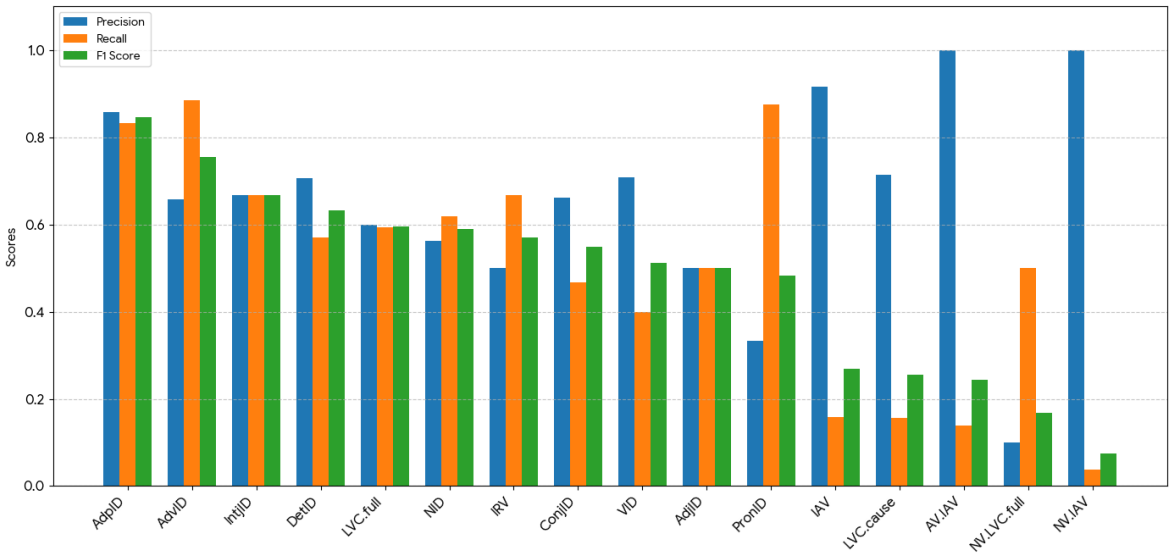

Figure 3: Precision, Recall, and F1-score across Ukrainian MWE classes.

rozraxovuvaty+Acc ‘to calculate smth.’ vs. *rozraxovuvaty na*+Acc ‘to count on smth.’, *zvil’nytysja* ‘to become vacant’ vs. *zvil’nytysja vid*+Gen ‘to get rid of smth.’.

Moreover, in the Ukrainian material we can identify polysemous verbs that occur both without a prepositional complement and with different prepositional complements in different meanings, cf.: *tjažity* ‘to gravitate’, *tjažity nad*+Ins ‘to weigh upon smth.’, *tjažity do*+Gen ‘to gravitate towards smth.’

Finally, the theoretically well-developed concepts of valency-predicted obligatory position and valency-predicted optional position (Vykhovanets’, 1993; Zahnitko, 2001) proved difficult to differentiate in practice. Firstly, this is due to the possibility of ellipsis – the omission of certain structural components in a sentence, including prepositions, which can be easily recovered from context, cf.: *Firmy* ***konkurujut’****, ščob prodaty teplo v merežu.* ‘Companies **are competing** for the right to sell heat to the network.’ (i.e. *Firmy* ***konkurujut’ odna z odnoju*** ‘Companies **compete with one another**’ = *Firmy* ***konkurujut’ miž soboju*** ‘Companies **compete among themselves**’). Secondly, in colloquial speech and social media posts, verbs may be used without their prepositional complements in non-normative way, cf.: *To* ***zaležyt’ vid*** *bahat’ox čynnykiv.* ‘It **depends on** many factors.’ (normative) vs. *To* ***zaležyt’****.* ‘It **depends**’ (colloquial). *To* ***zaležyt’ vid*** *toho, jak pytaty.* ‘It **depends on** how you ask.’ (normative) vs. *To* ***zaležyt’****, jak pytaty.* (Twitter 2017; colloquial, prepositional complement omitted).

5.3 Adposition idiom (AdpID)

A distinctive feature of Ukrainian MWEs is the significant number of secondary multi-component (compound) prepositions (Adposition idiom (AdpID)): ***u mežax*** *spravy* ’**within the scope of** proceedings’, ***za pidsumkamy*** *vizytu* ‘**based on the results of** the visit’. This is the third most common class of MWEs, comprising 865 units (see Figure 2). Such units reflect the processes of grammaticalization and phraseologization that are ongoing in the current stage of Ukrainian language development. For more information on the expansion of the group of secondary prepositions in Ukrainian, see (Zahnitko et al., 2007; Sytar and Zahnitko, 2025).

Similar processes within the prepositional and conjunctional subsystems of Ukrainian are noteworthy, and one can conclude that these units compete, cf.: ***nezvažajučy na*** *partiju* ‘**regardless of** party’ (preposition) vs. ***nezvažajučy na*** *te, xto do jakoï partiï naležyt’* ‘**regardless of** who belongs to which party’ (conjunction); ***nezaležno vid*** *rivnja osvity* ‘**regardless of** education level’ (preposition) vs. ***nezaležno vid*** *toho, jakyj riven’ osvity vin maje* ‘**regardless of** what education level he has’ (conjunction); ***vidpovidno do*** *real’nyx doxodiv* ‘**according to** actual income’ (preposition) vs. ***vidpovidno do*** *toho, jaki ïx real’ni doxody* ‘**according to** what their actual income is’ (conjunction), etc.

5.4 Variants of Idioms

Despite the fact that one of the characteristics of MWEs is their fixedness or limited flexibility, variants of idioms that are easily identified by native speakers and do not alter the holistic meaning of MWEs have proven to be characteristic of contemporary Ukrainian. However, this very variability of MWEs can complicate their automatic identification in text and the performance of other NLP tasks. According to our observations, idiom variants arise through the following transformations:

a) constructions with zero copula, typical of Ukrainian and other East Slavic languages: *Istyna zavždy poseredyni.* 'The truth always **lies** in the middle.' *U straxu velyki oči.* 'Fear **has** big eyes.' We classified such cases as verbal MWEs despite the formal absence of the verb.

b) introduction of additional components into MWEs: verbal phrase *prolyty svitlo* 'to shed light' modified by the adverbial particle *troxy* 'a little': *U rozmovi vin **prolyv troxy svitla** na perspektyvy ukraïns'kyx bankiv* 'In conversation, he **shed some light** on the prospects for Ukrainian banks.' The verbal phrase *povernutysja na Olimp* 'return to Olympus' is modified by the introduction of the possessive adjective *kyïvs'kyj*: *povernutysja na kyïvs'kyj Olimp* 'return to the Kyiv Olympus'. Common nominal idioms *krok upered* 'step forward' and *krok nazad* 'step back': *Te, ščo my robymo, - **krok upered, try vbik, potim odyn nazad**.* 'What we are doing is one **step forward, three steps to the side, then one step back**.'

c) replacement of components: the biblical expression *prodaty za mysku sočevyčnoï jušky* 'to sell (something) for a bowl of lentil stew' is transformed into *prodaty za tarilku boršču* 'to sell (something) for a plate of borshch' (a traditional Ukrainian dish): *Such propaganda exploits the servile mentality of the "nostalgic Soviet type", who is willing to sell freedom for a plate of borshch (Ukrainian Week, 2014; our transl. from Ukr.).*

d) omission of components: In the verbal phrase *zaxyščaty čest' [svoho] mundyra* 'to defend the honor of [one's] uniform', the noun for 'honor' is omitted in the interview text: *Vony* [pracivnyky sylovyx struktur] *duže **zaxyščajut' svij mundyr**.* 'They [law enforcement officers] strongly **defend their uniform**'.

Special attention was required during annotation for cases in which multiple types of MWEs were combined: *Amerykans'ka delehacija **vxodyt' do skladu** Parlaments'koï asambleï* 'The American delegation **forms part of** the Parliamentary Assembly.': *vxodyt' do*+Gen 'to be part of smth.' is an inherently adpositional verb, and *do skladu*+Gen is an adposition idiom, 'in smth.', lit. 'into the composition of smth.'. *Centr protydiï teroryzmu ta hibrydnym zahrozam **sprjamovuje** svoï **zusyllja na** vidbyttja kiberatak* 'The Center for Countering Terrorism and Hybrid Threats **directs its efforts toward** repelling cyberattacks.': *sprjamovuje zusyllja* 'directs efforts' is a light verb constructions, and *sprjamovuje na* 'directs toward' is an inherently adpositional verb.

These observations highlight the complexity of MWE phenomena in Ukrainian and point to areas requiring further investigation and annotation improvement. Currently, the annotation scheme does not provide a mechanism to link the variants as alternative forms of the same multiword expression. In future work, it would be beneficial to incorporate this functionality into the annotation framework.

6 Conclusions and Future Plans

The obtained results show an imbalance in the distribution of MWE types in Ukrainian interview texts. On the one hand, this imbalance highlights specific features of Ukrainian phraseology and grammar. On the other hand, these findings require further validation on a larger corpus and through the inclusion of data from other text styles.

The analysis of Ukrainian data also indicates the need to refine the existing MWE classification. In particular, we propose introducing a separate category within functional MWEs, namely Particle Idioms.

As a direction for future research, we plan to identify and analyze cases involving overlaps between different MWE types, which will contribute to a more precise and comprehensive description of multiword expressions in Ukrainian.

These findings and future research directions are enabled by the creation of a dedicated Ukrainian MWE resource. With the expansion of PARSEME in 2025 to include additional MWE types and languages, Ukrainian became part of the shared task for the first time. Supported by the UniDive project[12], we successfully integrated Ukrainian into this international initiative. Based on established annotation guidelines and previous linguistic research, the resulting resource represents the first

[12] https://unidive.lisn.upsaclay.fr/

comprehensive corpus of Ukrainian multiword expressions and supports both theoretical research and computational modeling.

Limitations

The presented resource has several limitations that should be mentioned. First, the size of the corpus is still relatively small compared to MWE datasets for Romanian, Hebrew, and Polish languages (more than 13,000 MWEs). Although it is useful for initial research, some rare constructions may not be well represented. In the future, the corpus should be expanded to include more text types and topics.

Second, even though the annotation scheme follows the PARSEME Shared Task 2.0 guidelines, some Ukrainian-specific constructions do not fit perfectly into the existing categories. In such cases, annotators had to rely on internal decisions, which may lead to small inconsistencies or unclear borderline cases.

Third, the automatic linguistic annotation produced by UDPipe 2 (such as tokenisation, lemmas, POS tags, or dependencies) may contain errors. While MWEs were annotated manually and independently of these layers, such automatic mistakes can still influence how some expressions are interpreted.

Finally, the annotation was carried out by a small group of annotators. Although they worked together and discussed difficult cases to ensure consistent decisions, involving more annotators and calculating formal inter-annotator agreement in future work would further increase the reliability of the resource.

These limitations point to several directions for improvement, such as extending the corpus, refining annotation rules, and adding more quality-control procedures.

Acknowledgments

We would like to thank the reviewers for their time and effort in reviewing this manuscript. We sincerely appreciate their valuable comments and suggestions, which greatly helped us improve the quality of the work. This research was partially funded by the Alexander von Humboldt Foundation, and this work received support from the COST Action CA21167 'UniDive'[13] (European Cooperation in Science and Technology). The authors are also grateful to Friedrich Schiller University Jena for providing the research facilities and support that made this work possible.

[13]https://unidive.lisn.upsaclay.fr/

References

Mykola Alefirenko. 1988. *Teoretychni pytannia frazeolohii*. Vyshcha shkola, Kharkiv.

Vira Bilonozhenko and 1 others. 1993. *Frazeolohichnyi slovnyk ukrainskoi movy*. Naukova dumka, Kyiv.

Vira Bilonozhenko and 1 others. 2003. *Slovnyk frazeolohizmiv ukrainskoi movy*. Naukova dumka, Kyiv.

Anita Braxatorisová, editor. 2024. *Synsemantika in Phrasem-Konstruktionen im Deutschen und anderen Sprachen*. Logos Verlag Berlin GmbH, Berlin.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Dmitrij Dobrovol'skij. 2011. Phraseologie und konstruktionsgrammatik. In Alexander Lasch and Alexander Ziem, editors, *Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze*, pages 110–130. Tübingen.

Voula Giouli and Verginica Barbu Mititelu, editors. 2024. *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*. Number 6 in Phraseology and Multiword Expressions. Language Science Press, Berlin. Published under CC BY 4.0.

Kateryna Horodens'ka. 2007. *Hramatychnyi slovnyk ukrainskoi movy: spoluchnyky*. Vydavnytstvo KhDU, Kherson.

Larysa Kolibaba and Valentyna Fursa. 2025. *Slovnyk diieslivnoho keruvannia: u 2 t.* Pidruchnyky i posibnyky, Kyiv.

Natalija Kushch. 2008. *Pryimennykova ekvivalentnist v ukrainskii hramatytsi: struktura, semantyka, funktsii*. Ph.D. thesis, Donetsk.

Alla Luchyk. 2006. *Slovnyk ekvivalentiv slova ukrainskoi movy*. Wydawnictwo Uniwersytetu Śląskiego, Katowice.

Tetiana Masyts'ka. 1998. *Hramatychna struktura dieslivnoi valentnosti*. RVV "Vezha" VDU im. Lesi Ukrainky, Lutsk.

Matvij Nomys. 1993. *Ukrainski prykazky, pryslivia i take inshe*. Lybid, Kyiv.

Karel Pala, Lukáš Svoboda, and Pavel Šmerk. 2008. Czech MWE database. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Agata Savary and Jakub Waszczuk. 2020. Polish corpus of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 32–43, online. Association for Computational Linguistics.

Maria Shvedova. 2020. The general regionally annotated corpus of ukrainian (grac, uacorpus.org): Architecture and functionality. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020), Volume I: Main Conference*, CEUR Workshop Proceedings, pages 489–506, Lviv, Ukraine.

Maria Shvedova, Arsenii Lukashevskyi, and Andriy Rysin. 2025. Developing a Universal Dependencies treebank for Ukrainian parliamentary speech. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 55–63, Vienna, Austria (online). Association for Computational Linguistics.

Ivelina Stoyanova, Svetlozara Leseva, and Maria Todorova. 2016. Towards the automatic identification of light verb constructions in bulgarian. In *Proceedings of CLIB 2016*, pages 28–37, Sofia, Bulgaria.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Hanna Sytar. 2010. Opysovi predykaty z modalnym komponentom v ukrainskii movi: struktura y semantyka. *Linhvistychni studii*, (20):145–151.

Hanna Sytar. 2017. *Syntaksychni frazeolohizmy v rozrizi konstruktsiinoi hramatyky*. TOV "Nilan-LTD", Vinnytsia. Monograph.

Hanna Sytar and Anatolii Zahnitko. 2025. Sekundární předložky s variantní valencí vyjadřující determinací akce v ukrajinštině. *Prace Filologiczne*, 80:325–350.

Viktor Uzhchenko and Dmytro Uzhchenko. 2005. *Frazeolohiia suchasnoi ukrainskoi movy*. Almamater, Luhansk.

Ivan Vykhovanets'. 1980. *Pryimennykova systema ukrainskoi movy*. Naukova dumka, Kyiv.

Ivan Vykhovanets'. 1988. *Chastyny movy v semantykohramatychnomu aspekti*. Naukova dumka, Kyiv.

Ivan Vykhovanets'. 1993. *Hramatyka ukrainskoi movy. Syntaksys*. Lybid, Kyiv.

Ivan Vykhovanets', Kateryna Horodens'ka, Anatolii Zahnitko, and Svitlana Sokolova. 2017. *Hramatyka suchasnoi ukrainskoi literaturnoi movy. Morfolohiia*. Vydavnychyi dim Dmytra Buraho, Kyiv.

Anatolii Zahnitko. 1996. *Teoretychna hramatyka ukrainskoi movy. Morfolohiia*. DonDU, Donetsk.

Anatolii Zahnitko. 2001. *Teoretychna hramatyka ukrainskoi movy. Syntaksys*. DonNU, Donetsk.

Anatolii Zahnitko, Illya Danyliuk, Hanna Sytar, and Inna Shchukina. 2007. *Slovnyk ukrainskykh pryimennykiv*. TOV VKF "BAO", Donetsk.

Anatolii Zahnitko and Anna Karataieva. 2012. *Slovnyk chastok: materialy i statti*. DonNU, Donetsk.

Anatolii Zahnitko, Kateryna Vynohradova, Illya Danyliuk, Nadija Zahnitko, Natalija Kushch, Maryna Orans'ka, Tetjana Kitaieva, Hanna Sytar, Valerija Chekalina, and Inna Shchukina. 2009. *Funktsionalno-komunikatyvna i tekstova paradyhma ukrainskykh pryimennykiv ta yikhnikh ekvivalentiv*. Weber (Donetska filiia), Donetsk.