PAWUK: extensive annotated web corpus of Ukrainian

Witold Kieraś[0000−0002−8062−5881][1], Łukasz Kobyliński[0000−0003−2462−0020][1],
Dorota Komosińska[0000−0002−2611−1214][1], Michał Rudolf[0000−0002−3115−9087][1],
Maria Shvedova[0000−0002−0759−1689][2], Aleksandra Zwierzchowska[0000−0002−7322−7535][1]

[1]Institute of Computer Science, Polish Academy of Sciences
{w.kieras, l.kobylinski, d.komosinska, m.rudolf,
a.zwierzchowska}@ipipan.waw.pl
[2]National Technical University "Kharkiv Polytechnic Institute", University of Jena
Mariia.Shvedova@khpi.edu.ua

Abstract. In this paper, we present PAWUK, the Polish Automatic Web corpus of UKrainian language. It is a linguistic corpus containing Ukrainian texts acquired from the Internet (selected web pages and social media accounts) and has been updated daily since 2022. It is automatically annotated with morphosyntactic tags, syntactic dependencies and named entities using Stanza framework with a model custom-built for Ukrainian to produce both Universal Dependencies tags and VESUM morphological tags. Users can interact with the corpus through a publicly available web interface.

Keywords: language corpus · Ukrainian language · natural language processing · PAWUK · regional variation

1 Introduction

Comprehensive balanced and representative linguistic corpora are a primary resource for studying language in its various registers. Building such a multi-purpose resource is both time- and cost-consuming and requires regular updating. On the other hand, the Internet has been a large source of text data for many years now and the idea of using these data as linguistic corpora goes back at least to the WaCky corpus [1]. Such corpora proved to be useful resources in various linguistic research in which a large dataset representing the most recent vocabulary is more important than its representativeness, as it is in the case of studying neologisms and neosemantisms. Recently such web-based corpora are also being used as the training data for building Large Language Models.

In this paper we present PAWUK[1], a corpus of contemporary Ukrainian built from various resources collected from the Internet and automatically updated on a daily basis. Our goal was to provide a reliable source of language data enriched

[1] https://pawuk.ipipan.waw.pl

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_7

by various layers of linguistic annotation: lemmatization, morphosyntactic annotation, dependency parsing and named entities recognition. Also we intended to provide the researchers with a relatively familiar environment at least with respect to the morphosyntactic annotation. Apart from the Universal Dependencies annotation scheme [7] we provide a morphosyntactic annotation consistent with the VESUM morphological dictionary of Ukrainian [12], allowing the user to query the corpus with the same tagset that was used in GRAC, the large reference corpus of contemporary Ukrainian[2]. To combine the Universal Dependencies annotation scheme and VESUM morphosyntactic scheme in a single natural language processing pipeline, a custom model of Stanza framework [9] needed to be trained. PAWUK's data sources were also manually assigned region labels according to the same scheme that was used in GRAC.

2 Related Work

PAWUK stands out as a significant online corpus of the Ukrainian language, joining resources such as GRAC, Zvidusil, UberText, and the Ukrainian components within multilingual collections like Aranea, ParlaMint, and the Leipzig Corpora Collection, as well as UkTenTen and Ukrainian Trends on Sketch Engine. However, PAWUK possesses key distinctions that make it a unique resource.

Notably, PAWUK adopts the regional annotation scheme pioneered by the large Ukrainian language corpus GRAC. Originally designed for studying regional variations in Ukrainian, GRAC evolved into a universal resource and a cornerstone of Ukrainian linguistics. GRAC.v.18 comprises texts primarily from printed sources, spanning from 1816 to the present, totaling 1.9 billion tokens. It selectively includes contemporary online media that exemplify standard literary Ukrainian. PAWUK serves as a valuable complement to GRAC by offering a substantially larger volume of contemporary material, encompassing both regional and non-standard language. The shared annotation principles enable researchers to extend linguistic investigations initially based on GRAC using PAWUK. A primary limitation is the partial implementation of regional annotation in PAWUK, currently applied to online media and some Telegram channels, while regional information is unavailable for Twitter and YouTube comments.

The majority of large web-based Ukrainian language corpora are static, containing texts downloaded from the Internet at a specific point in time. While some undergo periodic updates, such as UkTenTen (updated in 2014, 2020, and 2022), Araneum Ucrainicum (2014, 2015, 2021, and 2022), ParlaMint (twice in 2023), and the Ukrainian corpus from the Leipzig Corpora Collection (2011, 2014, 2019 onwards), they lack the capacity to observe daily language evolution in real time, a feature offered by monitoring corpora.

PAWUK and Ukrainian Trends function as monitoring corpora, automatically collecting new Ukrainian texts daily and providing searchable online interfaces. Both have been actively uploading texts since spring 2022. In comparison

[2] https://uacorpus.org/

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_7

to Ukrainian Trends, PAWUK is updated more frequently (twice a week/daily), downloads a greater volume of text, and includes social media content alongside standard news. This social media component offers valuable material for studying social media discourse and allows for the quicker identification of new words prevalent in online communication and memes. Furthermore, the social media data within PAWUK provides valuable insights into non-standard linguistic features and their regional distribution. Conversely, Ukrainian Trends, accessible through Sketch Engine, boasts a more sophisticated search interface, including the functionality to track trends for individual words.

For a more comprehensive list of current Ukrainian corpora, see Table 1.

3 Architecture of PAWUK

While designing the architecture of PAWUK, we had to address the challenge of being able to regularly collect large amounts of text from various sources on the Internet and to process them efficiently. More specifically, we had to tackle the issue of creating processing pipelines which could be configured independently for various text sources, solve such problems as proper language recognition, separation of main content from downloaded documents, text tagging, efficient storage and indexing.

Consequently, the main processing module is designed to manage the processes of acquiring, processing, and sending data for indexing from various online sources. The system architecture is built using Apache Airflow[3], which defines processing pipelines for different data sources (channels). The data acquisition process varies depending on the source.

In the case of websites, data is collected based on a list of URLs and specially prepared templates for each site. These templates include rules for parsing HTML, such as which elements to skip. In case of social networks data is gathered using lists of user profiles, with content retrieved through APIs provided by each platform.

Once acquired, the data goes through several processing stages:

- text conversion (extracting text from the source format, removing unnecessary characters, eliminating documents that are too short),
- language verification (for Twitter data, relying on source identification, for other sources, automatic verification using tools like Stanza),
- duplicate elimination (removing texts already present in the database),
- annotation (performing morphological and syntactic analysis, as detailed in section 5),
- indexing (using MTAS [2] for text indexing, making the indexed data available through a web service).

The workflow is constantly monitored and statistics are gathered and stored to signal any issues during various stages of the process, such as data collection,

[3] https://airflow.apache.org/

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_7

Table 1. Selected existing Ukrainian corpora.

Corpus	Size (in tokens)	Composition and annotation	Access
General Regionally Annotated Corpus of Ukrainian (GRAC)	1.9B	Various texts (1816-2024), VESUM-based annotation	Searchable
Ukrainian Brown corpus	1M	Balanced, various texts, partly manually annotated, VESUM-based annotation	Downloadable
UD Ukrainian IU Treebank	122K	Various texts, manually annotated corpus	Searchable, downloadable
UD Ukrainian ParlaMint Treebank	51K	Parliamentary transcripts, manually annotated corpus	Searchable, downloadable
Ukrainian ParlaMint	51M	Parliamentary transcripts (2002-2023), contains Ukrainian-Russian code switching; language annotation, UD pos annotation	Searchable, downloadable
UberText 2.0	3.3B	News, Wikipedia, social, fiction, court	Downloadable
Ukrainian Corpus (Leipzig University)	18B	News, web, Wikipedia (2011, 2014, 2019, 2020, 2021, 2022, 2023, 2024), without lemmatization	Searchable, samples of up to 1 million words for download
Ukrainian Web 2022 (ukTenTen22)	7.5B	Web texts (2014, 2020, 2022), MULTEXT-East Ukrainian pos annotation, UD pos annotation	Searchable
Ukrainian Trends	>1B	News, Wikipedia (since spring 2022), MULTEXT-East Ukrainian pos ann.	Searchable
Zvidusil (Laboratory of Ukrainian)	3B	Web texts (2018), UD annotation	Searchable
Araneum Ucrainicum	125M—1.25B	Web texts (2014, 2015, 2021, 2022), pos annotation	Searchable, registration is required
ParaRook Parallel corpora with German, English, French, Spanish, Chinese, Japanese, Persian	28M	Fiction	Searchable
Laboratory of Ukrainian Parallel corpora with English, Polish, French, German, Spanish, Portuguese	5M	Fiction	Searchable
Parallel Russian-Ukrainian	9M	Fiction, journalism	Searchable

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_7

tagging, or indexing. For example, we trigger notifications when the amount of data collected for particular sources drops below a set average value.

The overall architecture is presented in Figure 1.

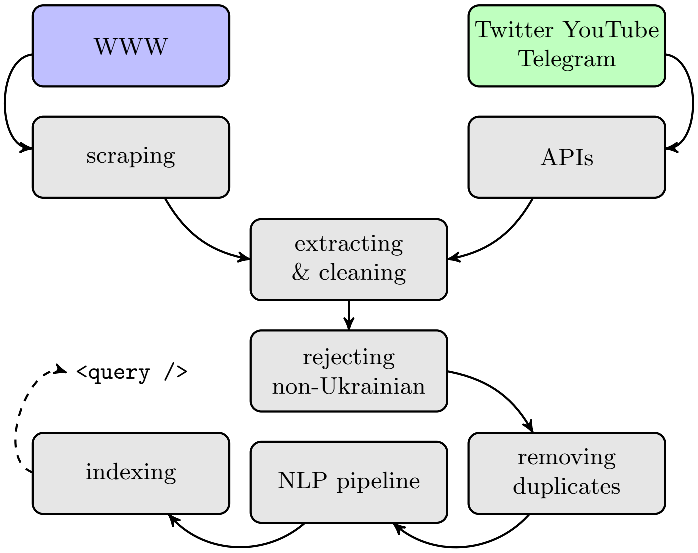


Fig. 1. Architecture of PAWUK.

4 Data Sources and Acquisition

PAWUK monitors 506 selected Ukrainian webpages as well as Ukrainian Twitter[4] (469 accounts), YouTube (294 channels) and Telegram (633 channels). Lists of news sites and Telegram channels were manually selected for the corpus.

The collection includes samples from news websites across all 24 regions of Ukraine as well as Kyiv, and features media outlets published in Ukrainian from other countries. When selecting news sites we aimed to create the most complete, representative, thematically diverse collection, which would include texts from all regions of Ukraine, both large cities and smaller towns. 366 out of 506 news websites are marked by regions.

The principles of selection of Telegram channels are the same as those for news sites. 200 out of 633 channels are marked by regions. There are official channels of institutions and public figures, as well as popular channels covering different topics: music, literature, local city channels.

[4] Data retrieval was discontinued on June 29, 2023, due to Twitter/X's decision to end the free API access.

DOI: 10.1007/978-3-031-97629-2_7

In the case of Twitter, data was collected from two sources — selected user accounts (targeting 469 accounts) and trends (most popular tweets from Ukraine). The account database initially consisted of several dozen manually selected users and was then automatically expanded based on their connections with other profiles. Accounts were periodically verified and updated.

The list of monitored YouTube accounts is created entirely automatically based on the list of most popular videos from Ukraine. Currently, it includes 294 accounts which are periodically verified. Only the text data from comments under the videos are collected.

Each document (a web article or a single post) is assigned metadata that can be used in queries. These consist of: title, author, region, URL, channel, subchannel and date of publication.

Data collection for the corpus occurs daily. Each channel is managed by a separate processing path defined in Apache Airflow. In case of any problems, the system administrator can respond and retry the process at any stage.

The starting point of the corpus collection, March 2022, was at the beginning of the active stage of Russia's war against Ukraine. Each text in the corpus is time-stamped with the day it was created. Thus, the corpus is an accurate testimony of the changes that occurred in the Ukrainian language during the war. This not only concerns the emergence of new words and meanings, but also, for example, a characteristic orthographic feature consisting in beginning proper names associated with Russia with a lowercase letter, also when starting a sentence or headline. On social media, users sometimes employ Unicode subscript characters to render the initial letter of a word visually smaller, as in “$_\text{р}$осія”.

5 Annotation

The corpus is automatically enriched with linguistic annotation accessible by the user through a corpus search engine. For the annotation process we use Stanza pipeline with a custom model for Ukrainian tweaked to produce VESUM morphosyntactic tags as XPOS values. The tagging and parsing models are trained using the sum of two independent resources:

- standard Ukrainian UD treebank[5] [7] converted to contain VESUM tags instead of existing MULTEXT-East tags as XPOS,
- BrUK[6] corpus manually annotated with VESUM tags (so called Ukrainian Brown corpus) parsed using the standard Stanza model.

The idea behind creating such a heterogeneous training set was to maximize the accuracy of XPOS tagging without impairing the accuracy of parsing as well as to diversify the training set as the Ukrainian UD treebank is rather small. The researchers of Ukrainian are more familiar with the VESUM tagset which

[5] The treebank was developed by the Institute of Ukrainian and is available at https://github.com/UniversalDependencies/UD_Ukrainian-IU/

[6] https://github.com/brown-uk

is more fine-grained (contains over 1,700 unique tags) and complies with the annotations of Ukrainian large reference corpus GRAC.

Table 2 presents evaluation results for the custom model. We provide the values of standard metrics that are commonly used in evaluation of NLP tagging and parsing frameworks. F1 measure for lemmatization and morphological features (Lem, UPOS, XPOS, UFeats) is provided for testing subsets of both datasets that were used in the training process. For parsing accuracy metrics (UAS, LAS, CLAS, MLAS, BLEX) only the UD treebank evaluation is provided as the other dataset does not contain a manual dependency annotation layer. The model is publicly available at https://github.com/ipipan/stanza-uk-pawuk.

For named entities annotation the standard Stanza model[7] was used.

Table 2. Evaluation results for testing subsets of the two resources used for training the custom model.

Eval. dataset	Lem	UPOS	XPOS	UFeats	AllTags	UAS	LAS	CLAS	MLAS	BLEX
BRuK	98.49	97.29	91.87	92.87	88.92	-	-	-	-	-
UD treebank	98.57	98.96	93.92	95.96	91.30	89.77	87.99	84.71	84.23	84.71
Total:	98.56	98.66	93.55	95.40	90.87	-	-	-	-	-

6 Interacting with the corpus

The interface of the corpus is very simple and basic. The front page contains the query field through which the user may immediately start working with the corpus. A simple query language cheat sheet may be found in the About subpage.

The simplest corpus query consists of a single word form typed in by the user which should result with the list of all concordances (words within context) containing this word. The multilayer linguistic annotation accessible through Corpus Query Language however allows for much more precise queries regarding lemmatization, morphology, syntax and named entities. The primary annotation scheme used in the corpus is the one provided by the UD scheme which recently became a *de facto* standard for morphosyntactic and dependency annotation. It consists of parts of speech (e.g. noun, adjective, numeral), morphological features, that is morphological categories (e.g. number, case, aspect) and their values (e.g. singular, accusative, imperfective respectively), and the dependency annotation, that is assigning dependency relations between words which are in a direct grammatical relation of certain type such as subject, object, adjunct, modifier etc. Apart from that, the linguistic annotation also includes the morphological tag of every word expressed in the tagset native to a given language (so called XPOS) which usually contains even more precise morphological information and is more familiar to the researchers working with the language on a

[7] https://stanfordnlp.github.io/stanza/ner_models.html

daily basis. In the case of PAWUK and Ukrainian, the native tagset is the one of VESUM morphological dictionary which was also used in GRAC.

The following examples give an overview of possible corpus queries from basic to more advanced ones, illustrating the expressive power of the linguistic annotation as well as the query language.

- `[orth="павуки"]` finds all occurrences of a given orthographic word.
- `[lemma="павук"]` finds all occurrences of a given lemma regardless of its inflectional form.
- `[upos="NOUN"]` finds all occurrences of words belonging to a given part of speech (as defined in the UD scheme), here: all nouns.
- `[ufeat="fem"]` finds all occurrences of words which were assigned a given value of a specific grammatical category, here: all words assigned the feminine gender regardless of their part of speech.
- `[deprel="nsubj"]` finds all occurrences of words that are dependents in a given grammatical relation, here: words that were assigned as nominal subject in the grammatical structure of a sentence.
- `[head.lemma="павук"]`, `[head.upos="ADJ"]`, `[head.ufeat="fem"]` work analogously to `lemma`, `upos`, `ufeat` mentioned above but they apply to syntactic head of a given word instead of the word itself.
- `[xpos="noun:anim:p:v_naz"]` finds all words assigned a given XPOS tag, in this case: animated nouns in plural number and nominative case.
- `<ne="PERS" />` finds all occurrences of named entities assigned one of the four categories: person (PERS, as in this query), organization (ORG), place (LOC), other (MISC).

The crucial part of the query language is that all the constraints described above can be combined in one query. For example, the following query will find all occurrences of the word павук ('spider') used as a subject in a sentence and occurring within a proper name referring to a person (see Figure 2):

```
[lemma="павук" & deprel="nsubj"] within <ne="PERS" />
```

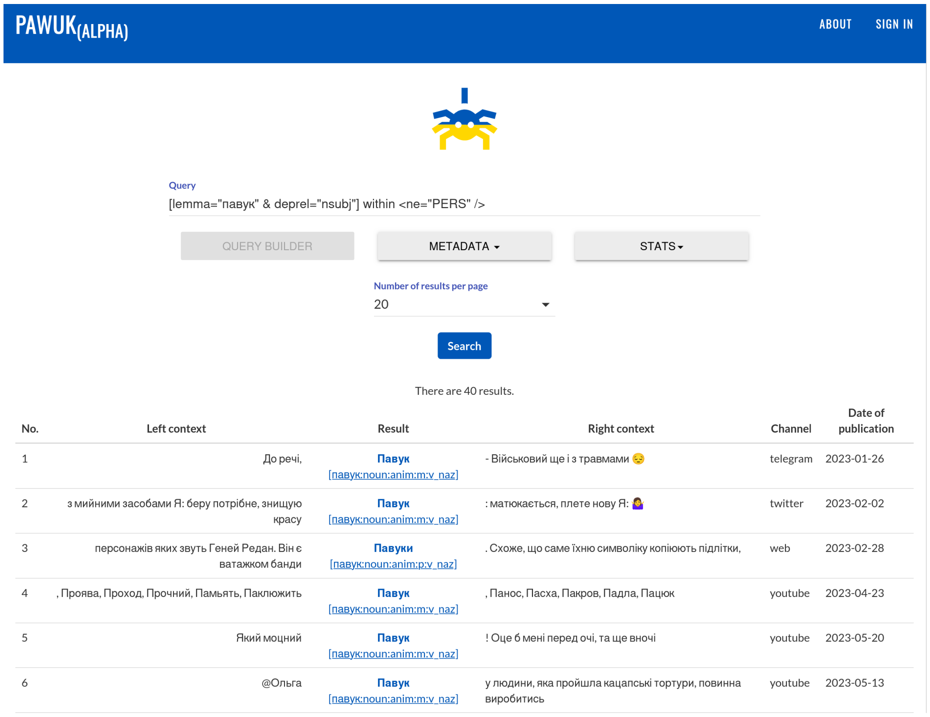
Among the results there could be an instance of Людина-Павук 'Spiderman'.

Additionally the annotation includes an `oov` (*out of vocabulary*) attribute (used as `[oov="true"]` in corpus queries) which allows to filter only those interpretations which do not comply with the VESUM dictionary. This feature proved to be useful in researching non-standard grammatical variants. For example, the query `[lemma="вухо" & oov="true"]` finds word вухи (non-standard variant of the nominative plural form, cf. вуха), which is not a literary norm but is still common in social networks.

The corpus queries may be restricted based on basic metadata, which allows to obtain results only from sources meeting certain specific criteria such as text title, text author, publication date and, most notably, the region of the country from which the source data originates. This is especially useful in researching regional variations of the language.

A graphical query builder allowing the user to construct queries without remembering the CQL syntax and names of all attributes will be available soon.

The multilayer linguistic annotation is a powerful tool for researchers interested in finding rare instances of words or grammatical constructions as well as (ir)regularities in the use of certain types of linguistic phenomena. It is worth noting that users may as well create their own corpus in the Korpusomat application [10] which uses the exact same annotation scheme and provides similar user interface to the one used in PAWUK, which makes the two tools complementary and useful for various linguistics research of modern Ukrainian.



Query
[lemma="павук" & deprel="nsubj"] within <ne="PERS" />

Number of results per page
20

There are 40 results.

No.	Left context	Result	Right context	Channel	Date of publication
1	До речі,	**Павук** [павук:noun:anim:m:v_naz]	- Військовий ще і з травмами 😔	telegram	2023-01-26
2	з мийними засобами Я: беру потрібне, знищую красу	**Павук** [павук:noun:anim:m:v_naz]	: матюкається, плете нову Я: 🤷‍♀️	twitter	2023-02-02
3	персонажів яких звуть Геней Редан. Він є ватажком банди	**Павуки** [павук:noun:anim:p:v_naz]	. Схоже, що саме їхню символіку копіюють підлітки,	web	2023-02-28
4	, Проява, Проход, Прочний, Памьять, Паклюжить	**Павук** [павук:noun:anim:m:v_naz]	, Панос, Пасха, Пакров, Падла, Пацюк	youtube	2023-04-23
5	Який моцний	**Павук** [павук:noun:anim:m:v_naz]	! Оце б мені перед очі, та ще вночі	youtube	2023-05-20
6	@Ольга	**Павук** [павук:noun:anim:m:v_naz]	у людини, яка пройшла кацапські тортури, повинна виробитись	youtube	2023-05-13

Fig. 2. The screenshot presents search results of the query example.

7 Use-case Scenario

As a case study of corpus-based research, we analyse herein the distribution of the variant prepositions *vid* and *od*, which share the semantic value 'from' and are historically variant forms of the same preposition. These two variants correlate with distinct Ukrainian dialectal groups. In the South-Western dialectal zone, the variant *vid* prevails, whereas *od* is the predominant form in the Northern dialects. The South-Eastern dialect area represents a more recent and heterogeneous formation, shaped by successive waves of migration and lacking

DOI: 10.1007/978-3-031-97629-2_7

clearly defined dialectal boundaries. Within this mosaic-like region, both variants — *vid* and *od* — are attested with varying frequency, as illustrated in map 269 of the Atlas of the Ukrainian Language, vol. 1 [8]. In the 20th century, the prepositional variant *vid* became established as the main standard form in the literary language, while *od* was largely relegated to spoken language and fictional texts [4].

We used PAWUK to investigate the geographical and quantitative persistence of the older variant preposition *od* in online texts.

The study revealed that the majority of occurrences of the variant *od* within the corpus originate from social media texts. The relative frequency of *od* across different text types is as follows: 3.2 occurrences per 1,000 combined uses of *od* and *vid* in YouTube comments, 1.6 in Twitter, 1.0 in Telegram, and 0.3 in web texts.

To ascertain whether a correlation exists between the distribution of the variant *od* in online texts from various regions and the dialectal areas where this variant is present, we examined web news articles and Telegram channels. While these sources yielded fewer instances of *od* compared to social networks, only this subset of the corpus was annotated for regional origin.

Given that the administrative oblasts represented in the corpus do not precisely align with dialectal distribution areas, we categorized the regions into two groups: those encompassing areas where *od* is used (regions that at least partially include areas of Northern or South-Eastern dialectal distribution) and those where this variant is absent. We excluded all Kyiv-based channels from the dataset due to the capital's diverse population, which renders the linguistic data non-indicative of the region's specific language use. Additionally, we excluded the prominent Lviv portal Zbruch, as it attracts contributions from individuals across Ukraine. Consequently, this analysis yielded a limited number of examples (see Table 3).

Table 3. Distribution of the variants *od* and *vid* in regions with and without *od* in dialects.

Dialectal *od*	*od*	*vid*
Yes	197	558,157
No	21	250,904

The corpus data nonetheless reveal a statistically significant correlation between the regional distribution of the preposition *od* and dialectal areas where this form is traditionally attested. Despite the infrequent overall occurrence of *od* within the corpus, its relative frequency is notably higher in regions with documented dialectal usage. A Fisher's exact test ($p < 0.000001$) and a chi-squared test ($\chi^2 = 45.57, p < 0.00000001$) corroborate that this association is not attributable to random variation.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_7

While the corpus of Internet texts does not constitute dialectal material per se and may not invariably reflect the direct diffusion of a word or feature within a region (as it could appear, for example, in a quotation or a metalinguistic context describing another speaker's language), the influence of local vernacular on written Internet texts is evident.

8 Summary and Future Work

Since the release of its initial version in 2023, PAWUK has been used in researching Ukrainian neologisms [6,5], usage of non-standard language varieties (such as surzhyk) in online communication [3] and teaching corpus linguistics [11].

The corpus is updated daily (approximately 1.5–2.3 million words every day; 2 million on average). Currently it consists of nearly 2 billion tokens (see Table 4).

Table 4. Corpus statistics.

Channel	Documents	Sentences	Tokens	Avg. per day
web	3.9M	61.3M	1.1B	1.3M
telegram	14.3M	32.6M	372M	310k
twitter	3.7M	7.5M	89M	200k
youtube	16.8M	38.4M	399M	360k
Total:	38.7M	139.8M	1.96B	2.17M

Although the process of collecting, annotating and indexing daily batches of data is purely automatic, the corpus needs some technical maintenance as some sources of data become unavailable (discontinued websites, social media data policy changes etc.) and new sources need to be added. Therefore, the primary goal is to maintain the corpus updating process and ensure its stability and reliability for users. In the future, we plan to add new data sources as they become available (e.g. new social networks with available APIs and new web sources). We also plan to improve the web interface to make interacting with the corpus more intuitive and user-friendly.

Acknowledgments. The work was partially supported by the Dariah.lab project, as well as the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN - Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19. The work of Maria Shvedova was partially supported by the visiting program of the Polish Academy of Sciences.

We would like to thank everyone who contributed to the collection of channels, including the students of the National Technical University "Kharkiv Polytechnic Institute" - Kateryna Astafieva, Kseniia Luzhna, Daria Lebedenko, Maryna Vozikova, Polina Poltavchenko, Arsenii Lukashevskyi, and Ivan Ahaiev, and participants of the Corpus Linguistics online course (as part of the digital philology program for Ukrainian students, supported by the University of Jena and the DAAD Foundation) - Nazar Kotsur, Nataliia Sheremet, Olha Tochylina, Yurii Petrov, Mariana Martyniv.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation **43**(3), 209–226 (2009). https://doi.org/10.1007/s10579-009-9081-4
2. Brouwer, M., Brugman, H., Kemps-Snijders, M.: MTAS: A Solr/Lucene based Multi Tier Annotation Search solution. In: Selected papers from the CLARIN Annual Conference 2016. Linköping Electronic Conference Proceedings (2017)
3. Dyka, L., Shvedova, M.: Synonymy in the ukrainian language and the problems of surzhyk (case study of synonyms with the meaning 'probably') (in Ukrainian). Mova: klasyčne - moderne - postmoderne (9), 50–71 (2023), https://ekmair.ukma.edu.ua/server/api/core/bitstreams/5bd65ed1-25c2-4abf-9403-0776e4c6cea1/content
4. Dyka, L., Shvedova, M.: History and Normative Status of the Preposition / Prefix од- in Modern Ukrainian (in Ukrainian). Slavia Orientalis **vol. LXXI**(No 4), 797–818 (2022). https://doi.org/10.24425/slo.2022.143220, http://journals.pan.pl/Content/125858/PDF/2022-04-SOR-06.pdf
5. Horoxova, T., Bojko, M.: Dynamics of lexical composition of the ukrainian language in the war period in 2022–2023 (in Ukrainian). Aktual'ni pytannja humanitarnyx nauk: miżvuzivs'kyj zbirnyk naukovyx prac' molodyx včenyx Drohobyc'koho deržavnoho pedahohičnoho universytetu imeni Ivana Franka **67**(1), 213–218 (2023), https://elibrary.kubg.edu.ua/id/eprint/46953/
6. Klym, J.: New invective vocabulary with the component "zet-" in modern internet communication: Based on the web corpus (in Ukrainian). In: Movnyj prostir sučasnoho svitu: tezy dopovidej VII Vseukraïns'koï naukovoï konferenciï studentiv, aspirantiv i molodyx učenyx. pp. 90–94. NaUKMA, Kyiv (2023), https://ekmair.ukma.edu.ua/items/747f8e6f-36c8-4f20-86dd-94ef0788476a
7. de Marneffe, M.C., Manning, C.D., Nivre, J., Zeman, D.: Universal Dependencies. Computational Linguistics **47**(2), 255–308 (07 2021), https://doi.org/10.1162/coli_a_00402
8. Matvijas, I. (ed.): Atlas of the Ukrainian Language, vol. 1. Naukova dumka, Kyiv (1984)
9. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020), https://nlp.stanford.edu/pubs/qi2020stanza.pdf
10. Saputa, K., Tomaszewska, A., Zawadzka-Paluektau, N., Kieraś, W., Kobyliński, L.: Korpusomat.eu: A Multilingual Platform for Building and Analysing Linguistic Corpora. In: Mikyška, J., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M. (eds.) Computational Science – ICCS 2023. 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part II. pp. 230–237. No. 14074 in Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2023). https://doi.org/https://doi.org/10.1007/978-3-031-36021-3_22, https://link.springer.com/chapter/10.1007/978-3-031-36021-3_22

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_7

11. Shvedova, M., Pospekhova, A.: Corpus linguistics course for philology students (in Ukrainian). In: Zbirnyk naukovyx prac' I Mižnarodnoï naukovoï konferenciï «Innovacijni texnolohiï v linhvistyci ta perekladi». pp. 60–64. L'viv (2024), https://books.ldubgd.edu.ua/index.php/m/catalog/book/203
12. Starko, V., Rysin, A.: VESUM: A large morphological dictionary of Ukrainian as a dynamic tool. In: Computational Linguistics and Intelligent Systems. vol. 6th Int. Conf, pp. 71–80. COLINS, Gliwice (2022), https://ceur-ws.org/Vol-3171/paper8.pdf

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_7