# A Contemporary News Corpus of Ukrainian (CNC-UA): Compilation, Annotation, Publication

**Stefan Fischer, Kateryna Haidarzhyi, Jörg Knappen, Olha Polishchuk, Yuliya Stodolinska, Elke Teich**
Universität des Saarlandes
Campus A2.2, 66123 Saarbrücken, Germany
{stefan.fischer, kateryna.haidarzhyi, olha.polishchuk, yuliya.stodolinska}@uni-saarland.de
{j.knappen, e.teich}@mx.uni-saarland.de

**Abstract**

We present a corpus of contemporary Ukrainian news articles published between 2019 and 2022 on the news website of the national public broadcaster of Ukraine, commonly known as SUSPILNE. The current release comprises 87 210 364 words in 292 955 texts. Texts are annotated with titles and their time of publication. In addition, the corpus has been linguistically annotated at the token level with a dependency parser. To provide further aspects for investigation, a topic model was trained on the corpus. The corpus is hosted (Fischer et al., 2023) at the Saarbrücken CLARIN center under a CC BY-NC-ND 4.0 license and available in two tab-separated formats: CoNLL-U (de Marneffe et al., 2021) and vertical text format (VRT) as used by the IMS Open Corpus Workbench (CWB; Evert and Hardie, 2011) and CQPweb (Hardie, 2012). We show examples of using the CQPweb interface, which allows to extract the quantitative data necessary for distributional and collocation analyses of the CNC-UA. As the CNC-UA contains news texts documenting recent events, it is highly relevant not only for linguistic analyses of the modern Ukrainian language but also for socio-cultural and political studies.

**Keywords:** corpus creation, contemporary news, Ukrainian

## 1. Introduction

This paper introduces a new contemporary news corpus for Ukrainian (CNC-UA), a corpus of modern Ukrainian news texts covering the 38-month period from November 2019 until December 2022. The corpus comprises 292 955 texts, mainly news articles but also reports with long tables. The CNC-UA is made available under a Creative Commons Attribution-Non-Commercial-NoDerivs 4.0 International License, while the underlying raw data are subject to the copyright of Суспільне Мовлення (Suspilne Movlennya, henceforth SUSPILNE), the Public Broadcasting Company of Ukraine.

While a number of corpora of Ukrainian do exist, overall the resource situation for Ukrainian is mixed. On the one hand, the availability of the existing corpora is often complicated (e.g. various search interfaces, no possibility of download, incomplete documentation). On the other hand, larger corpora are often not specialized enough to allow for serious linguistic or sociopolitical analysis (e.g. lack of contextual metadata).

Furthermore, due to the current situation in Ukraine, it is important to engage in the preservation and archival of news texts in a non-proprietary way for sociopolitical and linguistic documentation and subsequent scientific analysis. This is what motivated the creation of the CNC-UA.

This paper is structured as follows. We give an overview of existing Ukrainian corpora (Section 2) and explain the corpus building and annotation process of the CNC-UA (Section 3). To make the corpus as useful as possible, we have embedded it in an existing eco-system including a web-based corpus analysis platform as well as various standard formats. We describe the access to the CNC-UA and downloadable formats in Section 4. To demonstrate the application of the corpus and its accompanying infrastructure, we provide a short exploratory analysis (Section 5). We conclude with a brief summary and outlook in Sections 6 and 7.

## 2. Related Ukrainian Corpora

While Ukrainian can still be considered a low-resource language, the number of Ukrainian corpora is steadily increasing. Many of these corpora are available online, for example, the Ukrainian Language Corpus (Darchuk, 2017), the General Regionally Annotated Corpus of the Ukrainian Language (GRAC; Shvedova, 2020), the Ukrainian Text Corpus (Department of General and Applied Linguistics and Slavic Philology, Vasyl Stus Donetsk National University, 2023), and the Ukrainian Brown Corpus (BRUK; Starko and Rysin, 2023). Other important corpora, such as the National Ukrainian Linguistic Corpus (Shyrokov, 2011) or the Computer Fund for Innovation (Karpilovska, 2007) are currently inaccessible to the general public.

The General Regionally Annotated Corpus of the Ukrainian Language (GRAC) has a volume of



*The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024*, pages 1–7

1.781 billion tokens (v17). It is a vast and organized collection of texts in Ukrainian, allowing users to create subcorpora, search for words and grammatical forms, analyze search results, sort data, form balanced samples, and obtain statistical information via the Sketch Engine platform. The GRAC is a diachronic corpus spanning from 1816 to 2022 and contains over 130 000 texts from various genres. It contains a large subcorpus of journalism that includes collections of newspapers from the 19th and 20th centuries, contemporary newspapers, and texts from news sites on the web. The majority of texts come from printed sources. Notably, it includes a large corpus of diaspora texts, totaling about 40 million tokens. The corpus comprises both original and translated Ukrainian texts. However, no license is specified for this corpus and it is not downloadable.

The Ukrainian Language Corpus[1] (Darchuk, 2017) consists of morphological, syntactic, and semantic annotation layers. Currently, it contains more than 100 million tokens, partitioned into six subcorpora: journalism, fiction, scientific texts, legislative texts, poetic language, and folklore texts. The corpus is accessible through a corpus manager and is not downloadable.

The Ukrainian Text Corpus[2] (Department of General and Applied Linguistics and Slavic Philology, Vasyl Stus Donetsk National University, 2023) contains 120 000 word occurrences. It includes various genres such as journalistic, fictional, scientific, legislative, poetic, and folklore texts that have been processed automatically at morpheme, word, phrase, and sentence levels (part-of-speech, grammatical form, syntactic function).

The Ukrainian Brown Corpus (BRUK; Starko and Rysin, 2023) is an ongoing project aiming at creating an open, genre-balanced corpus of the modern Ukrainian language. The corpus contains text samples from 2010 to 2020 with a volume of 1 million words. It is built on the same principles as the well-known English Brown corpus. The texts were automatically tokenized, lemmatized and annotated with part-of-speech tags. The manual disambiguation of the corpus is still ongoing. Selected texts from national, regional, and local media, both print and online, make up approximately 25% of the BRUK. It is available for download under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

The University of Leipzig has collected a large corpus (Leipzig Corpora Collection, 2014; Goldhahn et al., 2012) of Ukrainian internet texts dating from 2014. This corpus is downloadable and it contains 1 546 330 404 tokens. Also, the online corpus portal allows to visualize text connectivity, and even offers a graph of interconnections.

The LORELEI Ukrainian Representative Language Pack (Tracey et al., 2020) includes Ukrainian monolingual texts, Ukrainian-English parallel and comparable texts, annotations, additional resources, and related software tools. This corpus contains 111 million words of Ukrainian text, of which about 700 000 words have been translated into English.

The UberText 2.0 corpus (Chaplynskyi, 2023) contains 3.274 billion tokens in 8.59 million texts. It has five subcorpora: news, fiction, social, wikipedia and court. The news subcorpus contains 2.173 billion tokens (before filtering), which are scraped from 38 central, regional, and industry-specific news websites. Among other steps, the processing pipeline includes lemmatization and POS tagging. The five subcorpora can be downloaded individually in different formats.

The Institute for Ukrainian (NGO)[3] is a joint Polish-Ukrainian project. The team has developed several corpora and a dedicated morphological analyzer. The Gold Standard Universal Dependencies Corpus for Ukrainian (Kotsyba et al., 2022) contains 140 000 tokens. The texts in the corpus have been manually annotated with morphological and syntactic dependency annotations. The corpus comprises a variety of text types, including articles, news, posts, textbooks, letters, fairy tales, and fiction. The news texts make up 5.6% of the total amount. There are no restrictions on the time of creation for the texts in the collection. The web corpus Zvidusil (Institute for Ukrainian, 2018), which is also part of this project, contains over 2.8 billion tokens. It has been automatically annotated and homonymy has been removed. The corpus includes texts from various freely available sources, such as user posts on social media, found mostly on the internet. To search the corpus[4], one can specify subcorpora based on source, title, author, and time of appearance of the texts. However, the texts do not contain further metadata. Statistical information and information about the search results are available. The web corpus Zvidusil includes news periodicals, such as Vysokyi Zamok, Den, Dzerkalo Tyzhnia, Zbruch, Radio Svoboda, Tyzhden, Ukraina Moloda, Ukrainska Pravda, etc.

The Ukrainian Web Corpus (ukTenTen 2022)[5] is a corpus composed of Ukrainian texts gathered from the internet. It belongs to the TenTen family (Suchomel, 2020; Jakubíček et al., 2013; Suchomel and Pomikálek, 2012) of corpora, which are a set of

[1] http://www.mova.info/corpus.aspx
[2] http://corpora.donnu.edu.ua/
[3] https://github.com/UniversalDependencies/UD_Ukrainian-IU
[4] https://mova.institute/bonito/run.cgi/corp_info?corpname=zvidusil
[5] https://www.sketchengine.eu/uktenten-ukrainian-corpus/

web corpora created using the same method with a target size of over 10 billion words. ukTenTen 2022 contains over 9.5 billion tokens and is classified by genre and topics. The data for the corpus consists of texts from May 2014, July–August 2020, and October–December 2023.

Another recent corpus is the Ukrainian parliamentary corpus ParlaMint-UA (Kopp et al., 2023), which contains plenary proceedings of the Rada and covers the period from May 2002 to November 2023. It is available in two versions: a collection of plain texts with TSV metadata of the plenary speeches and the collection of plenary speeches with added automatic linguistic annotations. ParlaMint-UA 4.0.1 has more than 51 million tokens, 41 million words, 3.4 million sentences, and 429 thousand statements from 2532 speakers in 1723 meetings.

This overview is not complete, e.g. Shvedova (2020) and Chaplynskyi (2023) also describe Ukrainian corpora not mentioned here. Although a number of research teams are currently working on the automated and manual creation and annotation of different Ukrainian language corpora, some aspects still require additional data and further enhancements. Currently, there is a need for data from contemporary sources such as news, which reflect the ongoing processes in the society and the current linguistic developments. Even though other corpus projects incorporate news data, e.g. GRAC, ukTenTen, UberText, Zvidusil and other corpora mentioned above, their texts are from various sources or time periods, and they are often limited due to copyright issues. Considering the current context, the creation of the CNC-UA is timely. Firstly, the CNC-UA was established in 2023 and covers news data from November 2019 to December 2022 with the potential of expansion. This period covers two significant events, not only in Ukraine but also in Europe and the world: the coronavirus epidemic and Russia's full-scale invasion of Ukraine. Secondly, the CNC-UA is based on news texts of SUSPILNE. This media platform presents international, national, and regional news on a wide range of topics, i.e. world, culture, sports, economy, politics, nature, etc. SUSPILNE is one of few independent media companies in Ukraine, which (in contrast to predominantly private media platforms) has had its unique role as a state-owned and authoritative representative of Ukrainian media.

The CNC-UA fills a gap by providing a middle ground between the large corpora of Ukrainian, e.g. GRAC or UberText 2.0, and smaller, hand-crafted corpora such as BRUK. Furthermore, it is based on official data from a single source and not based on web-scraping. It can be used for training and fine-tuning models for the Ukrainian language as well as sociopolitical, historical and linguistic studies.

# 3. Corpus Building and Annotation

## 3.1. Origin and Content of Texts

The first publicly available release of the CNC-UA covers three full years, namely from the end of 2019 until 2022. The corpus is based on raw data in SQL format received from SUSPILNE in December 2022, which forms the basis of its news[6] website. The contents of other media channels, i.e. Facebook, Telegram, YouTube, that also belong to SUSPILNE, are not represented in the corpus.

The raw texts were not labelled with topics, although the website of SUSPILNE uses an extensive tagging system for topics (e.g. crimes of Russian Federation, corruption, weapons, Crimea, Ukraine-EU, Ukraine and NATO) as well as categories (e.g. politics, economics, world, regions, people, technologies, nature, culture, sports). Due to the absence of the original topical annotation in the raw data, a model of eight topics was trained on the lemmatized texts (see Section 3.3). Interestingly, a small number of 34 English texts was identified with the fastText library (Joulin et al., 2016a,b).

## 3.2. Statistics

The CNC-UA contains 87 210 364 tokens in 292 955 texts in this first release. The breakdown of the amount of texts and tokens over time is shown in Table 1. The number of texts increases each year. Taking into consideration that the current version of the corpus contains texts from November 2019 to December 2022, the statistics do not represent the whole year of 2019. Nevertheless, they show that the number of accessible texts and tokens grows steadily.

| Year | # Texts | # Tokens |
|---|---|---|
| 2019 | 6887 | 1 813 880 |
| 2020 | 81 157 | 21 997 108 |
| 2021 | 95 974 | 30 275 296 |
| 2022 | 108 937 | 33 124 080 |

Table 1: Size of the CNC-UA over time.

## 3.3. Metadata and Annotation

At the initial stage, it was established that the received data contained the following per-text information: id, title, body, timestamp. The CNC-UA was then enriched with the following information for each text: hour, month, weekday, year, year_month. Linguistic annotations were then added using the Stanza NLP tools (v1.4.2; Qi et al., 2020), whose Ukrainian model was trained on data from the Universal Dependencies project (v2.8). For additional

[6] https://suspilne.media

metadata, a topic model of eight topics (administration, crime, culture, everyday, health, international, sports and war; see Table 2) was trained on the lemmatized texts with the MALLET toolkit (McCallum, 2002). For the development of the eight topics over time see Figure 1. These metadata can be useful for exploring the linguistic changes that occurred over time and studying patterns that can be traced by topic (see Section 5 below for an example).

| Topic | Keywords |
|---|---|
| Administration | head, job, hryvnia, council, work<br>голова, робота, гривня, рада, працювати |
| Crime | police, man, court, criminal, report<br>поліція, чоловік, суд, кримінальний, повідомити |
| Culture | museum, person, history, job, project<br>музей, людина, історія, робота, проєкт |
| Everyday | say, child, person, tell, talk<br>казати, дитина, людина, розповісти, говорити |
| Health | case, person, COVID, coronavirus, hospital<br>випадок, людина, covid, коронавірус, лікарня |
| International | Russia, president, country, Russian, report<br>росія, президент, країна, російський, повідомляти |
| Sports | match, team, championship, world, competition<br>матч, команда, чемпіонат, світ, змагання |
| War | military, Russian, report, shelling, territory<br>військовий, російський, повідомити, обстріл, територія |

Table 2: Topic labels and top-5 ranked keywords by topic in CNC-UA.

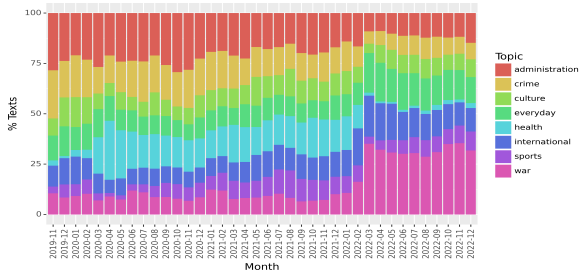


Figure 1: Dominant topics over time in CNC-UA.

Currently, tokens are annotated with the following linguistic information, which is provided by the Stanza NLP tools: word form, lemma, part-of-speech tags (universal and language-specific), morphological features (e.g. animacy, case, gender, number) and dependency information (head and relation type). The language-specific part-of-speech tags are based on the MULTEXT-East Morphosyntactic Specifications, Version 4.

## 4. Access and Download

The CNC-UA is designed and built according to the FAIR data principles (Wilkinson et al., 2016) and can be accessed from a research data repository specializing in linguistic corpora. It is hosted at the CLARIN-D repository[7] at Saarland University. The corpus is findable by a persistent and globally unique identifier (see Section 10). The CNC-UA is described by rich CMDI (Broeder et al., 2011) metadata with a link to the landing page of the corpus. The metadata are indexed and searchable by the CLARIN Virtual Language Observatory (Van Uytvanck et al., 2010, 2012). The CNC-UA is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license.

We provide files in several common formats. Besides, there is an option for exploring the corpus through a web-based corpus analysis platform. The CNC-UA can be downloaded in two tab-separated formats: CoNLL-U and VRT. The CoNLL-U format (de Marneffe et al., 2021) contains all linguistic annotations provided by the Stanza NLP tools. In particular, it allows one to work with dependency trees. We also provide the corpus in the vertical text format (VRT), which is of interest to (corpus) linguists as it allows them to encode the corpus on their own CWB (Evert and Hardie, 2011) or CQPweb (Hardie, 2012) servers. For users who do not need their own installation, we also provide a CQPweb server.[8] Lastly, the metadata is available in tabular format.

## 5. Exploratory Analysis

In order to explore the linguistic similarities and differences of the corpus data within different time periods and topics we can use the CQPweb interface. Using queries for concordances, distributional data, frequency lists, and collocations enables us to identify not only the variations in the linguistic contexts but also the various textual patterns within the existing corpus.

To demonstrate the potential of the corpus on the lexical level, we have chosen the concept of democracy. Specifically, we look at the noun демократія (en: democracy, translit: demokratiya), which exemplifies formal stylistically-marked political vocabulary. To analyze the representation of democracy, the query “[lemma="демократія"]” is used, which returns 1126 matches in 861 different texts. The distribution of hits for this query based on classification by *year* (Figure 2) demonstrates the fluctuation of occurrences over the 38 months period. The distribution of hits for this query based on classification by *topic* (Figure 3) shows that the majority of occurrences are within the topic *International*, followed by *Culture* and *Administration*.

The collocation analysis of the query “[lemma="демократія"]” (collocation window “1 to the left” and “1 to the right”, frequency at least 5) demonstrates that the noun демократія can be immediately linked to most open-class parts of speech, with adjectives having the highest mutual

[7] https://fedora.clarin-d.uni-saarland.de/

[8] http://corpora.clarin-d.uni-saarland.de/cqpweb/

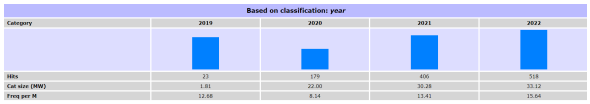


Figure 2: CQPweb: Distribution of hits for "[lemma="демократія"]" classified by year.

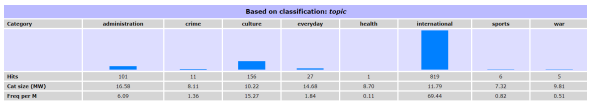


Figure 3: CQPweb: Distribution of hits for "[lemma="демократія"]" classified by topic.

information (MI) score. Ліберальний (en: liberal, translit: liberalnyi), having an MI score of 11.2, represents the strongest first-order collocate for the analyzed lemma. The noun with the highest MI score for "[lemma="демократія"]" is взірець (en: role model, translit: vzirets) with an MI score of 9.3. The only verb with a relatively high MI score, namely 5.0, is the verb захищати (en: protect, translit: zakhyshchaty). Further collocation analyses of "[lemma="демократія"]" might provide additional insights into the conceptualisation of демократія in Ukrainian news.

In this section we have demonstrated an example of using the CNC-UA with the CQPweb interface, which allows to extract the quantitative data necessary for distributional and collocation analyses. The concordances, distributional data, and frequency lists for the query "[lemma="демократія"]" show that the ongoing progress of democracy in Ukraine is reflected in the state-owned media.

## 6. Discussion and Future Work

The current size of the CNC-UA and the results of our first analyses are already promising. However, we acknowledge that the corpus in its current first release has certain limitations, which need to be taken into account and addressed in future work. In comparison to many other corpora, the corpus is not balanced, which is by design.

First of all, the time span and the size of the CNC-UA could be expanded. The current version contains the materials officially received from SUSPILNE in 2022 at the initial stage of our cooperation. The dataset covers the period starting from 2019 when the new official orthographic rules for Ukrainian were introduced thus reflecting the most recent changes in the Ukrainian language. The dataset goes up to 2022, encompassing a total number of 38 months, which is quite substantial for a news corpus based on a single source. Nevertheless, adding more recent data from 2023 and later to the CNC-UA will significantly increase its value and provide the most up-to-date news dataset for contemporary linguistic analysis and interdisciplinary studies.

In contrast to (smaller) manually annotated corpora, the processing of CNC-UA depends on the availability of external NLP tools for Ukrainian. While the performance of the Stanza pipeline was evaluated[9] on Universal Dependencies (UD) treebanks, an additional evaluation on the corpus is worthwhile. During our work with the corpus, no major problems were found. However, the lemmatization of non-Ukrainian proper names left room for improvement in some cases, e.g. *Scholz*. Also, more metadata at the text level would be desirable.

Lastly, our preliminary experiments using the CNC-UA have raised the issue that the texts contain links to related articles. As a result, the titles of articles are repeated throughout the corpus, which occasionally falsely raises the number of word occurrences and the distribution of search hits. This issue, i.e. boilerplate detection, may require additional cleaning or filtering of the dataset and should be addressed in our future work to increase the accuracy of results.

## 7. Summary and Conclusions

In this paper we have presented a new corpus of modern Ukrainian news texts. The first publicly available release covers the period from 2019 to 2022. We have placed the CNC-UA in the landscape of existing corpora of Ukrainian in order to demonstrate that it fills a gap by providing a middle ground between the existing large corpora of Ukrainian and the smaller, hand-crafted corpora. The corpus provides metadata at the text level and has been linguistically annotated at the token level with a dependency parser. The current release of the CNC-UA is open and available for download in several common formats. Besides that, we provide an option for exploring the corpus through a web-based corpus analysis platform. As the CNC-UA contains news texts documenting recent events, it is highly relevant for linguistic analysis, as well as sociopolitical, cultural, and interdisciplinary research.

## 8. Acknowledgements

This research is funded by Deutsche Forschungsgemeinschaft (DFG) – project IDs 232722074 (SFB 1102) and 460033370 (NFDI Text+).

[9] https://stanfordnlp.github.io/stanza/performance.html

## 9. Bibliographical References


Daan Broeder, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, and Andreas Witt. 2011. A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI). In *Proceedings of Balisage: The Markup Conference 2011*, volume 7 of *Balisage Series on Markup Technologies*, Montréal, Canada.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Nataliia Darchuk. 2017. Mozhlyvosti semantychnoyi rozmitky korpusu ukrainskoyi movy (KUM). *Naukovyi chasopys Natsionalnoho pedahohichnoho universytetu im. M.P. Drahomanova*, 9(15):18–28.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Department of General and Applied Linguistics and Slavic Philology, Vasyl Stus Donetsk National University. 2023. Korpusy tekstiv ukrainskoi movy. http://corpora.donnu.edu.ua/. Accessed: 2024-04-02.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*. University of Birmingham, UK.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul.

Andrew Hardie. 2012. CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Institute for Ukrainian. 2018. Zvidusil. https://mova.institute/bonito/run.cgi/corp_info?corpname=zvidusil. Accessed: 2024-04-02.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen corpus family. In *Proceedings of the 7th international Corpus Linguistics conference (CL2013)*, pages 125–127, Lancaster, UK. Lancaster University: UCREL.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Yevheniia Karpilovska. 2007. Tendentsii rozvytku suchasnoho ukrainskoho leksykonu: chynnyky stabilizatsii innovatsii. *Ukrainska mova*, 4:3–15.

Natalia Kotsyba, Bohdan Moskalevskyi, and Mykhailo Romanenko. 2022. Gold standard Universal Dependencies corpus for Ukrainian. https://github.com/UniversalDependencies/UD_Ukrainian-IU. Accessed: 2024-04-02.

Leipzig Corpora Collection. 2014. Ukrainian mixed corpus based on material from 2014. https://corpora.uni-leipzig.de?corpusId=ukr_mixed_2014. Accessed: 2024-04-02.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. https://mimno.github.io/Mallet/. Accessed: 2024-04-02.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. 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*, pages 101–108, 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)*, CEUR Workshop Proceedings, pages 489–506, Lviv, Ukraine. CEUR-WS.org.

Volodymyr Shyrokov. 2011. *Ukrainska leksykohrafiia v zahalnoslovianskomu konteksti: teoriia, praktyka, typolohiia*, chapter Zastosuvannia Ukrainskoho natsionalnoho linhvistychnoho korpusu v leksykohrafii ta linhvistychnykh ekspertyzakh. Vydavnychyi dim Dmytra Buraho.

Vasyl Starko and Andriy Rysin. 2023. Creating a POS gold standard corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural*

*Language Processing Workshop (UNLP)*, pages 91–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Vít Suchomel. 2020. *Better Web Corpora For Corpus Linguistics And NLP*. Ph.D. thesis, Masaryk University, Faculty of Informatics, Brno, Czech Republic.

Vít Suchomel and Jan Pomikálek. 2012. Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43.

Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the Virtual Language Observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1029–1034, Istanbul, Turkey. European Language Resources Association (ELRA).

Dieter Van Uytvanck, Claus Zinn, Daan Broeder, Peter Wittenburg, and Mariano Gardellini. 2010. Virtual Language Observatory: The portal to the language resources and technology universe. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 900–903, Valletta, Malta. European Language Resources Association (ELRA).

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

## 10. Language Resource References

Fischer, Stefan and Haidarzhyi, Kateryna and Knappen, Jörg and Stodolinska, Yuliya and Teich, Elke. 2023. *Contemporary News Corpus for Ukrainian (CNC-UA)*. CLARIND-UdS. PID http://hdl.handle.net/21.11119/0000-000E-1C5C-D.

Kopp, Matyáš and Kryvenko, Anna and Rii, Andriana. 2023. *Ukrainian parliamentary corpus ParlaMint-UA 4.0.1*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1900.

Tracey, Jennifer and Strassel, Stephanie and Graff, David and Wright, Jonathan and Chen, Song and Ryant, Neville and Ma, Xiaoyi and Kulick, Seth and Delgado, Dana and Arrigo, Michael. 2020. *LORELEI Ukrainian Representative Language Pack*. Linguistic Data Consortium, ISLRN 551-143-444-242-2.