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

Stefan Fischer, Kateryna Haidarzhyi, Jörg Knappen,
Olha Polishchuk, Yuliya Stodolinska, Elke Teich
Saarland University

UNLP 2024

UNIVERSITÄT
DES
SAARLANDES

Introduction

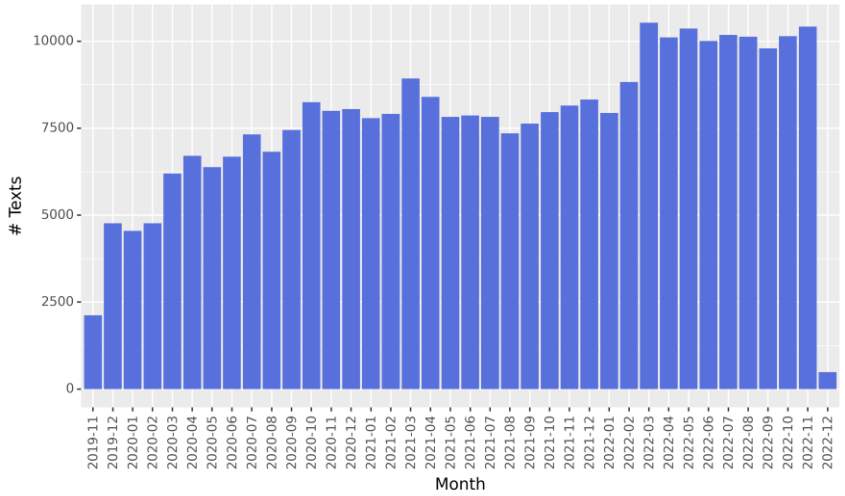
- Ukrainian is still a low-resource language
- Creation of many resources in the past years: BRUC, GRAC, KUM, KTUM, Leipzig Corpora, LORELEI, ParlaMint-UA, UberText, UD Treebank, ukTenTen, Zvidusil, etc.
- **C**ontemporary **N**ews **C**orpus for **Ukrainian**
- **CNC-UA** provides news from a single media platform

Contemporary News Corpus for Ukrainian (CNC-UA)

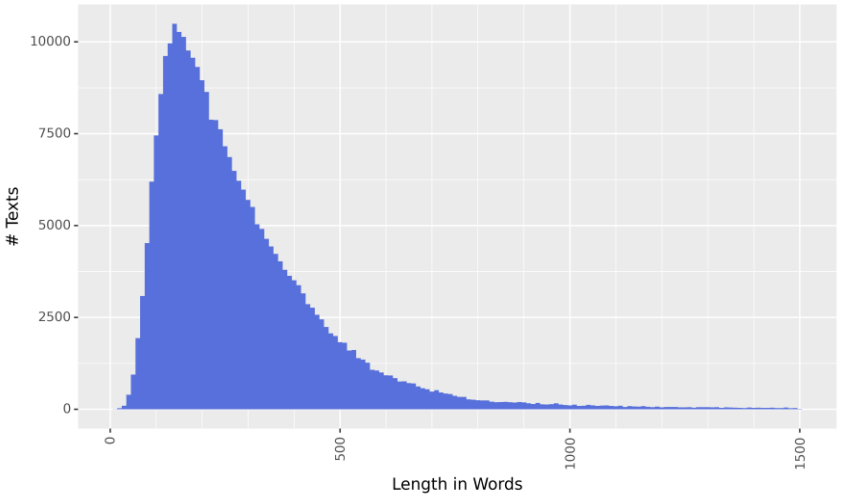
CNC-UA

- Source: modern Ukrainian news texts from SUSPILNE
- Coverage: November 2019 until December 2022
- Size: 292 955 articles
- Access: CC BY-NC-ND 4.0
- Use: for training and fine-tuning models for the Ukrainian language; sociopolitical, historical and linguistic studies

Statistics



Distribution of texts over time



Histogram of text lengths (avg. 298)

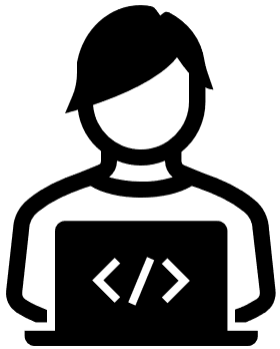
Source

- CNC-UA is based on data for the website of SUSPILNE https://suspilne.media/
- Other channels of SUSPILNE are not represented (Facebook, Telegram, YouTube)
- Data was received directly from SUSPILNE
  → No web scraping

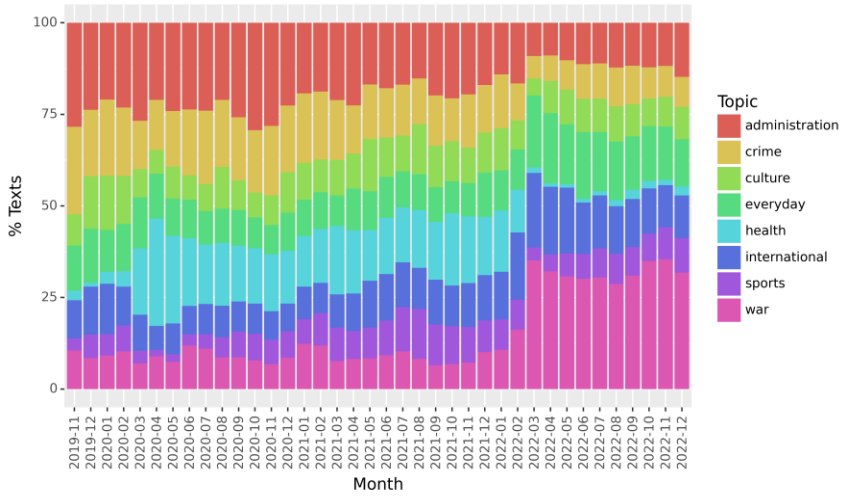
https://corp.suspilne.media/

Processing

- Linguistic annotation with Stanza NLP tools:
  Lemmas, parts of speech, morphology, dependencies
- Metadata: title, date and time from raw data
- No categorization information as on website
  → Enriched with topic model
- Language identification with fastText



Topic Model



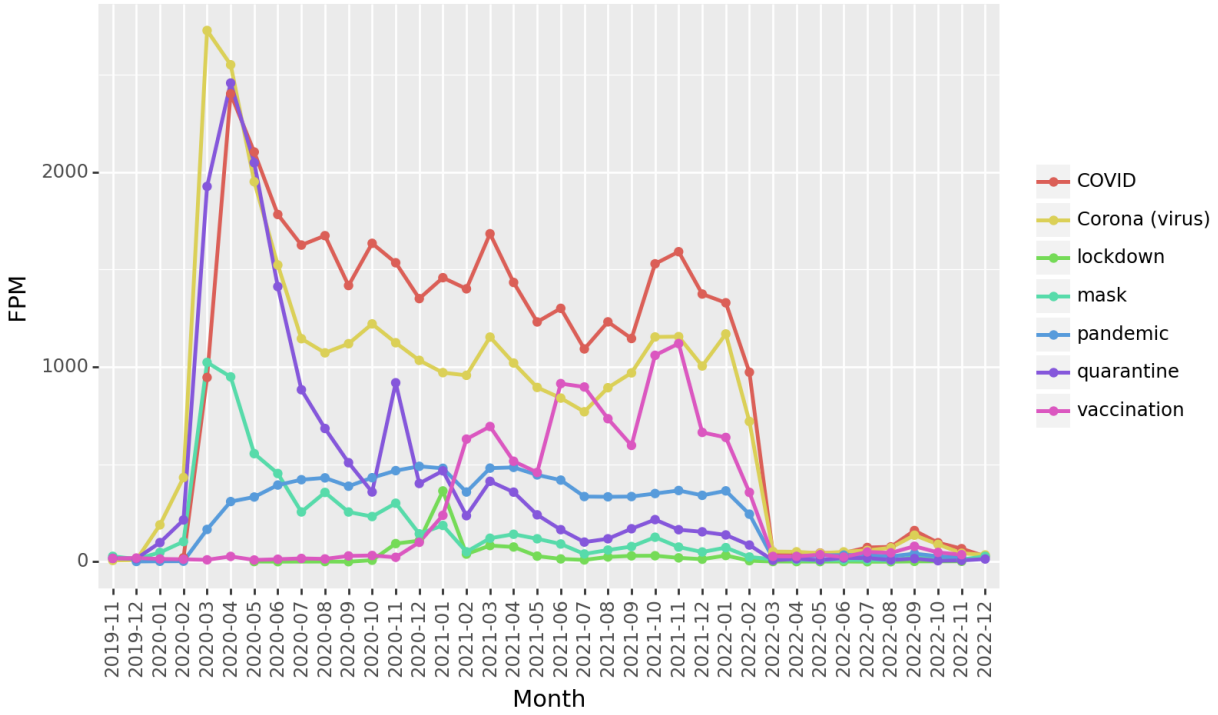
Diachronic usage of topics

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

Topic labels and keywords

- Topic modeling with MALLET
- 8 topics trained on lemmatized texts

Keyword Analysis



Frequency of keywords related to COVID-19

Access

- Creative Commons license (CC BY-NC-ND 4.0)
- CLARIND-UdS Language Resource Repository
  http://hdl.handle.net/21.11119/0000-000E-1C5C-D
  - CoNLL-U format
    (dependency analysis)
  - vertical text format
    (IMS Open Corpus Workbench and CQPweb)
- Installation of CQPweb corpus platform
  http://corpora.clarin-d.uni-saarland.de/cqpweb/

No	Text	Solution 1 to 50	Page 1 / 1,490	
1	2096	Розповідає голова офісу ВООЗ в Україні Імунітет проти	COVID - 19	запобігає повторному зараженню до 10 місяців – науковці
2	2096	науковці Як відсвяткувати Новий рік і не підхопити	COVID - 19	. Поради ЦГЗ На Дніпропетровщині затримали жителя Покрова
3	4525	святкування Як відсвяткувати Новий рік і не підхопити	COVID - 19	. Поради ЦГЗ Управління Верховного комісара ООН у
4	4548	році Як відсвяткувати Новий рік і не підхопити	COVID - 19	. Поради ЦГЗ Верховний суд Великої Британії відклав

CQPweb

No	Text	Solution 1 to 50	Page 1 / 1,490	
1	2096	Розповідає голова офісу ВООЗ в Україні Імунітет проти	COVID - 19	запобігає повторному зараженню до 10 місяців – науковці
2	2096	науковці Як відсвяткувати Новий рік і не підхопити	COVID - 19	. Поради ЦГЗ На Дніпропетровщині затримали жителя Покрова
3	4525	святкування Як відсвяткувати Новий рік і не підхопити	COVID - 19	. Поради ЦГЗ Управління Верховного комісара ООН у
4	4548	році Як відсвяткувати Новий рік і не підхопити	COVID - 19	. Поради ЦГЗ Верховний суд Великої Британії відклав
5	6137	Як відсвяткувати Новий рік і не підхопити	COVID - 19	. Поради ЦГЗ Як не зганьбитися перед колегами
6	14149	Офісу президента України Андрій Єрмак Поширення нового коронавірусу	COVID - 19	не вплине на розклад Олімпійських і Паралімпійських ігор
7	14149	45183 лабораторно підтверджених випадків нової коронавірусної інфекції (	COVID - 19	) , із них 1116 — летальні .
8	14379	провінція Хубей) , де зафіксований спалах коронавірусу	COVID - 19	, запланована на 18 лютого , повідомляється на
9	14379	року у світі зареєстровано 64437 лабораторно підтверджених випадків	COVID - 19	, з них 1383 — летальні . Що
10	15425	45183 лабораторно підтверджених випадків нової коронавірусної інфекції (	COVID - 19	) , із них 1116 — летальні .
11	15456	неправдиву інформацію про начебто лабораторно підтверджені випадки коронавірусу	COVID - 19	в Україні . " Ця інформація не відповідає
12	15494	приземлився в Україні . Інкубаційний період для появи	COVID - 19	у людей , які були на усіх бортах
13	15616	Київ - Москва " за результатами лабораторного дослідження	COVID - 19	не підтверджено . Нагадуємо : жодного випадку COVID
14	15616	19 не підтверджено . Нагадуємо : жодного випадку	COVID - 19	в Україні не зареєстровано . Раніше у прикордонній
15	15846	В Україні жодного лабораторно підтвердженого випадку коронавірусу	COVID - 19	не зафіксовано . У МОЗ уточнили , що
16	15846	до даних ВООЗ вірус , який спричиняє захворювання	COVID - 19	, називається SARS - CoV - 2 .
17	15912	що в Україні жодного лабораторно підтвердженого випадку коронавірусу	COVID - 19	не зафіксовано . Що відомо Третя людина померла
18	16064	1 725 випадків . В Україні жодного підтвердженого випадку	COVID - 19	не зафіксовано . Авторка — Ніна Булах ,
19	16104	Івано-Франківщині На Сумщині через негоду Пік епідемії вірусу	COVID - 19	вже минув і припав на період між 23
20	16298	треба робити в разі підозри на інфікування коронавірусом	COVID - 19	. Мультфільми опубліковані на сайті відомства . У
21	16373	інтернаті № 91 пройшов урок присвячений коронавірусу	COVID - 19	. Про це повідомляє кореспондентка Суспільного . Дітей
22	16802	" Якщо ви перебуваєте в зоні ризику	COVID - 19	(коронавірус) , будь ласка , залишайтеся
23	16901	Це додаткові заходи безпеки проти вірусу "	COVID - 19	" , або коронавірусу , який ширяться країнами
24	17126	вона знаходилась на карантині для обстеження на коронавірус	COVID - 19	. Про це повідомляє російський " Інтерфакс "
25	17234	Ми отримали діагностикум для визначення нового китайського коронавірусу	COVID - 19	. І за допомогою приладів , які ми
26	17337	метою мінімізації ризиків занесення на територію України збудника	COVID - 19	, відповідно до статей 33 та 96 Закону
27	17347	щодо проведення масових заходів через поширення коронавірусної інфекції	COVID - 19	" , - йдеться в повідомленні представництва України
28	17401	по країнах , в яких були зареєстровані випадки	COVID - 19	" , — зазначив головний санлікар . За
29	17458	валідатори не працюють . Спалах коронавірусу нового типу	COVID - 19	обернувся глобальною пандемією — інфекція , яка забрала
30	17458	які виділяються з носа або рота хворого	COVID - 19	при кашлі або чханні . Медичні експерти ВООЗ

Future Work

- Extension with more recent data
- Independent evaluation of annotation
- Detection and removal of boilerplate text

Conclusion

CNC-UA

- Contemporary news texts from SUSPILNE website
- Built with data from publisher, without web scraping
- Available under a Creative Commons license
- Highly relevant for linguistic analysis, as well as sociopolitical, cultural, and interdisciplinary research

Thank you!