The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality

Maria Shvedova[0000-0002-0759-1689]

Kyiv National Linguistic University

corpus.textiv@gmail.com

Abstract. The paper presents the General Regionally Annotated Corpus of Ukrainian, which is publicly available (GRAC: uacorpus.org), searchable online and counts more than 400 million tokens, representing most genres of written texts. It also features regional annotation, i. e. about 50 percent of the texts are attributed with regard to the different regions of Ukraine or countries of the diaspora. If the author is known, the text is linked to their home region(s). The journalistic texts are annotated with regard to the place where the edition is published. This feature differs the GRAC from a majority of general linguistic corpora.

Keywords: Ukrainian language, corpus, diachronic evolution, regional variation.

1 Introduction

Currently many major national languages have large universal corpora, known as “reference” corpora (cf. *das Deutsche Referenzkorpus* – DeReKo) or “national” corpora (a label dating back ultimately to the British National Corpus). These corpora are large, representative for different genres of written language, have a certain depth of (usually morphological and metatextual) annotation and can be used for many different linguistic purposes.

The Ukrainian language lacks a publicly available linguistic corpus. Still there is a need of a corpus in the present-day linguistics. Independently researchers compile different corpora of Ukrainian for separate research purposes with different size and functionality. As the community lacks a universal tool, a researcher may build their own corpus according to their needs. For example, the team UberText [30] has built a large Internet-based corpus which is published with shuffled sentences. This published version of corpus fits well the purpose of statistic analysis or sentence-level structures but cannot be used for studies of text structure or cohesion in discourse. The absence of a (reasonably) universal corpus for Ukrainian is still an issue.

The paper presents the GRAC (uacorpus.org) [25] – the General Regionally Annotated Corpus of Ukrainian, which is intended to fill in this gap. It is searchable online and counts more than 400 million tokens, representing most genres of written texts. About 50 percent of the texts are attributed with regard to the different regions of Ukraine or countries of the diaspora. This feature differs the GRAC from a majority of general linguistic corpora. This is due to the heavy regional and dialectal impact on the development of the norm(s) of Modern Standard Ukrainian at least until about 1950 and to certain extent until now (the problem that is *per se* a subject of further studies).

Our paper has the following structure. The second section discusses the existing corpora of Ukrainian. In the third section we will proceed to the contents and general architecture of the GRAC. The fourth section features the regional annotation. The fifth section discusses the metatextual annotation with regard to the genres, types of texts, dates and sources. The sixth section is dedicated to the translations that are also included into the corpus, whereas in the seventh and the eighth one we embark on the question of the morphological annotation and orthographic regime, which are known to be deeply interdependent parameters. In the section nine, the search engine and query are presented. The section ten briefly presents the results and future plans.

2 Principal Corpora of Ukrainian: an Overview

The only corpus of written Ukrainian where the dates of creation are specified for all the texts and these texts are annotated with regard to their region and genre is the General Regionally Annotated Corpus of Ukrainian (GRAC, http://uacorpus.org). Using the corpus one can study linguistic phenomena synchronically and diachronically according to the style and genre, as well as their statistical distribution with regard to the regions.

The corpus is developed by the present author in collaboration with Ruprecht von Waldenfels (Germany), Serhii Yaryhin (Ukraine), Andrii Rysin (USA), Vasyl Starko (Ukraine), Tymofii Nikolaienko (Ukraine), Mikhail Kruk (USA), Michał Woźniak (Poland).

2.1 The Ukrainian Text Corpus (KTUM)

This corpus is created in the Laboratory of computer linguistics, Philological Institute of the Taras Shevchenko Kyiv National University under direction of Natalia Darchuk since 2003 [9]. It counts 100 million tokens, including legal texts (1,6 million), academic texts (8,7 million), poetic texts (800 thousand tokens), journalism (40 million), fiction (36 million).

The corpus is available online [13]. The KTUM is the first publicly available Ukrainian corpus, searchable online since 2010. One can search it for a word, word form or grammatical features of a single word or of a two-term combination.

The main disadvantages of the corpus are the type of morphological markup that demands selection of a fully specified grammatical form whereas separate morphological features are not searchable (e. g. it is impossible to specify any part of speech in genitive unless a particular POS is selected). Texts are not annotated with regard to the date when they are created, only publication dates are provided instead. The technical base of the corpus is also to be renewed.

2.2 Parallel Ukrainian-Russian and Russian-Ukrainian Corpora within the Russian National Corpus

This corpus is compiled by Maria Shvedova, within the project of the Institute of the Russian Language, Russian Academy of Sciences, in 2009-2012. It counts 9,3 million tokens, original texts and translations from 1774 (Russian-language texts by Grigory Skovoroda) to 2011. The corpora are available for online search [17]. The Ukrainian-Russian part counts 6,5 million tokens: fiction (431 texts), journalism (29 texts), popular science (10 texts), legal (5 texts), letters of Ukrainian writers (180 texts). The Russian-Ukrainian part counts 2,8 million tokens: fiction (171), journalism (6 texts), popular science (6 texts), legal (6 texts). The texts are taken from printed source and from the web. Textual pairs (original and translation) are aligned sentence-by-sentence with the free program HunAlign [32]; later this alignment is manually corrected using the Euclid program [26].

The corpus is searchable by word, word form, grammatical feature or a set of features. Strings up to ten tokens each are searchable. The search results are represented as Russian-Ukrainian bilingual pairs with information about source. Within the concordance contexts can be expanded to three sentences.

2.3 The Corpus Project of the Laboratory of Ukrainian

It is compiled by Natalia Kotsyba, Bohdan Moskalevskyi and Mykhailo Romanenko [14]. Within the project several corpora with a dedicated morphological analyzer are developed, viz. a treebank with manually resolved homonymy and manual tagging (140 thousand tokens), the Zvidusil web corpus with automatic syntactic annotation (about 3 billion tokens), as well as parallel corpora that count the following size of foreign texts: Polish (4 million), English (1.5 million), French (0,5 million), German (190 thousand), Spanish (65 thousand), Portuguese (16 thousand). Morphological tagging in parallel corpora is made automatically according to the Universal Dependencies system. The corpora use the NoSketchEngine Platform and are available online for searching using the interfaces Bonito and KonText.

2.4 The Ukrainian Web Corpus of the Leipzig University (Germany)

The corpus counts 1,5 billion tokens, available for online search [3]. The corpus is built by means of webcrawling and contains only texts created before 2014 from the Internet (mostly news). There is no morphological tagging, only word forms are searchable. The corpus shows textual examples and collocations and plots graphs that visualize frequencies of word forms co-occurred in a sentence.

2.5 Corpus of Spoken Rusyn

The Corpus of Spoken Rusyn is compiled by Achim Rabus and Ruprecht von Waldenfels in Freiburg University in 2017 [6]. The recordings for the corpus are made in 2015 in Ukraine (Zakarpattia region), Slovakia, Poland and Hungary. The recordings are manually annotated, each recorded fragment is accompanied by the respective aligned transcript which is annotated and searchable. The search is made only by word form, a regional subcorpus can be customized. The fragments of recordings can be played and downloaded.

2.6 The Brown Corpus of Ukrainian

This name is given to a corpus compiled by Vasyl Starko, Andrii Rysin et al., after the well-known Brown corpus of English. It is a small-sized balanced corpus (1 million tokens) for building a statistical language model used for automatical language processing. It is currently under development [27].

2.7 The Ubertext Corpora

The corpora of Ukrainian texts: news, Wikipedia, fiction, web texts [30] are developed by the Ubertext team and available for downloading with shuffled sentences due to copyright reasons.

2.8 The Corpus of the Chtyvo Library

It counts about 600 million words [7]. It includes automatically recognized texts of the books from the Chtyvo electronic library without postprocessing. The search is made by exact query lacking lemmatization, morphological analysis and correction of OCR mistakes.

2.9 Summary

The main properties of the corpora described in the present section are summarized in the Table 1.

Table 1. Corpora of Ukrainian

Corpus	Size	Texts included	URL
GRAC	437 million	different genres	uacorpus.org
KTUM	100 million	journalism, fiction, academic, legal, poetic	mova.info/corpus.aspx
Parallel Ukr-Rus	9,3 million	fiction, journalism, academic, legal, letters	ruscorpora.ru/new/search-para-uk.html

Zvidusil	3 billion	Web texts	mova.institute
Laboratory of Ukrainian, treebank	140 thousand	different genres	mova.institute
Laboratory of Ukrainian, parallel	6 million	Fiction	mova.institute
Leipzig University	1,5 billion	Web texts	https://corpora.uni-leipzig.de/en?corpusId=ukr_mixed_2014
Spoken Rusyn	140 thousand	spoken texts	parasolcorpus.org/Varchola1
Brown	1 million	different genres	https://github.com/brown-uk/corpus
Ubertext	600 million	news, Wikipedia, fiction, Web texts	https://lang.org.ua/uk/corpora/
Chtyvo	600 million	books: fiction, academic texts, journalism	chtyvo.org.ua/search-korpus/

In this section we used some of the material cited in the paper [10]. Corpora that are not available online are not included into our overview. In particular, we do not know the actual size and functionality of the Ukrainian national linguistic corpus [31], developed by the Ukrainian language-information foundation of the Ukrainian national academy of sciences. It was published under restricted access and was not available to us.

3 Contents and Architecture of the GRAC

The GRAC is intended to encompass texts in (different forms of) the standard-oriented written Ukrainian since the first texts in Modern Ukrainian to the present day. This corpus is designed to enable a study of diachronic, diatopic and normative variation of the standard-oriented language. Examples of corpora-based studies are [24] (on possessive pronouns) or [22] (on syntactic variation).

The latest version (GRAC v.7.0) was created in December 2019. The chronological period covered by the corpus spans from 1816 to 2019. The size of the corpus is more than 430 million tokens in more than 45 thousand texts of different genres created by about 15 thousand individually known authors. The texts are either scanned and recognized from printed sources or taken (often with additional OCR and copyediting) from the Internet; the sites are listed on the website of the corpus. The corpus represents different genres, styles and regions. For all the texts the respective dates of creation are specified, many texts also have data concerning their publication. For the first time a subcorpus of Ukrainian diaspora is created (such texts had been largely neglected in the Ukrainian corpus linguistics, whereas they are crucial for studying normative variation, cf. [1] and [29]).

The corpus is currently designed to include prose, but poetic texts are also planned to be included into its future expansion. As of early 2020, the *Aeneid* by

Kotliarevskyi, the first masterpiece of Modern Ukrainian, was the only poetic text to be featured in the Corpus. The corpus currently has no specific annotation for poetic texts (metrical structure etc.)

Translations are also included into the corpus as they have played a significant role in the development of the standard language.

About 50% of the texts belong to the domain of fiction.

Among non-fiction a large subcorpus of journalism is to be pointed out that includes collections of newspapers of 1888-1893 (*Dilo, Ruslan, Narod, Červona Ukrajina, Bukovyna, Narodna Časopys'*), 1905 (*Xliborob*), 1913-1918 (*Dilo, Ruslan, Djilo i Nove Slovo, Vil'na Ukrajina, Vistnyk Sojuza vyzvolennja Ukrajiny, Krakivs'ki visty, L'vivs'ki visti, Vistnyk polityky, literatury i žyttja*), 1919-1943 (*Strilec', Šljax do voli, Visti VUCVK, Dilo, Meta, Novyj čas, Svoboda, Ukrajins'kyj Beskyd, Vil'na Ukrajina, Červona Ukrajina, Krakivs'ki visty, Červonyj Peremyšl', L'vivs'ki visti, Ukrajins'ki ščodenni visty, Ukrajins'kyj Visnyk, Holos Pidkarpattja*), contemporary newspapers from different regions (*Ukrajina moloda, Vysokyj zamok, Slovo, Visnyk SNAU, Kryms'ka svitlycja, Naš den', Čornomorec', Nyva, Šalom Alejxem, Vinnyčyna, Svoboda, Učytel', Visnyk odes'koji advokatury, Visti Donbasu, Vpered, Krajeznavstvo Zaporižžja, Licejist, Nasha hazeta [Novodnistrovs'k], Smiljanochka, Spivdružnist', 21-j kanal, 7 dniv, Volyns'ki novyny, Vorskla* et al.) as well as texts from news sites in the web (as such editions sometimes use machine translation for translating news, only sites that have the Ukrainian version as the single one were used as sources). Another large subcorpus consists of academic and educational texts: monographs, dissertations, scholarly papers, textbooks. There exists a separate subcorpus of religious texts, including among others two Ukrainian translations of the Bible. The subcorpus of ego texts also features memoires, letters and diaries, including a considerable corpus of Facebook posts representing blogs of people from all the Ukrainian regions and from the diaspora. Also included are subcorpora of spoken genres, viz. speeches and interviews.

The GRAC includes also some dictionaries featuring phrasal examples and idioms, among others Dictionary of Ukrainian by B. Hrinchenko and Russian-Ukrainian Dictionary of Idioms by I. Vyrhan and M. Pylynska. Using the corpus instruments the dictionaries can be searched not only by lexemes but also by lexico-grammatical patterns used in the examples and cited idioms.

The majority of texts come from printed sources. There are also smaller subcorpora of Internet texts (news, Facebook), visual media texts (translations and subtitles for movies and TV shows) and family documents (correspondence and diaries).

The family letters (about 800 texts) are collected by the students of the Lviv Polytechnical University advised by Professor Olena Levchenko. The texts are transcribed; of all the texts included into the GRAC, they depart in the largest extent from the standard-oriented language. The original pictures of these letters are also available within the corpus.

4 Regional Annotation

Worldwide, there exist different large corpora of regional and/or national linguistic variations. Examples include corpora of New World Englishes (including the corpus [12], developed since 1988). The corpus of Global Web-based English [5] has about 2 billion tokens. The regional and/or national linguistic variants are discussed e. g. in [21] or in the more recent [15]. Corpora-based research of lexicon and grammatical categories in different regional English varieties is conducted, eg of the Perfect gram (in the volume [33]).

A similar corpus system is built by Mark Davies for Spanish [4]. The corpus includes 2 billion tokens from 21 Spanish-speaking country and the United States.

Within the Russian National corpus a subcorpus of foreign press is including featuring a collection of Hrodna Region newspapers in Russian and Belarusian. There exist also corpus-based research of lexical and grammatical characteristics [20]. A corpus of post-1991 Russian texts of Ukraine is also created and some pilot studies performed [23].

Russian linguists have studied the Russian language with regard to the regional variety using the bulk of the Internet texts and the functionality of searching by regions (especially in blogs) of the Yandex and Google search engines [18]. These search tools are not a dedicated linguistic tool lacking capacities for linguistic search and do not yield the exact quantitative information necessary for a statistical research. Nevertheless using the Internet for searching regionally marked linguistic data has evident advantages such as large textual database, geographical and stylistical diversity, speed and easiness of search, sometimes presence of regional, chronological and authorship information.

These corpora and research are designed for pluricentric languages, that is the ones that function in different states where local linguistic norms, and, ultimately, local language variations are formed. Modern Ukrainian is not pluricentric *sensu stricto*, although it was formed historically in different centres and many local differences (between the East, the West and the diaspora) are still present today [11], [16] . Hence, in the research of the variation of the Ukrainian language several approaches and methods evolved in the study of pluricentric languages can be applied [8].

The regional markup of the corpus is based on the contemporary administrative structure of Ukraine. These boundaries as a pure convention that does not suggest any correlation between the regional (standard-oriented) variation and the dialectal boundaries. The texts of all the *oblasts* of Ukraine and from Crimea are represented in the corpus. The regions are grouped into six macroregions. Nearly a half of the texts in the corpora have regional markup.

Texts belong to the region were the author (or the translator, for a translated text) was born, studied and/or lived for more than ten year. The media texts are marked by the region where the respective media appeared. A single text can belong to different regional subcorpora (if the author or the translator was born, studied or lived for a long time in different regions). Alongside with regional subcorpora, there are subcorpora of diaspora (the United States, Canada, Poland, Germany, the United Kingdom, France etc.).

5 Metatextual Annotation

All the texts in the corpus are annotated by the year when they were written or by the last year when the text could be possibly created. The translated texts are marked by the year when the translation was made. The date of creation is the main date of a given text within the corpus. According to this date the text falls into the respective chronological subcorpus and counts in statistical research. Additionally, the date of the edition, used in the corpus, can be specified.

The corpus features information (if available) on the authors: year of birth, gender, and region(s), where the author was born, studied at a university and/or lived for more for than ten years.

The corpus contains four types of media: newspapers, magazines, TV channels and news sites. At the search page users can specify either the name of an edition or its type.

Each edition has information on the region where it appears (or used to appear). Since the information on the authors in media is often inaccessible, the regional affiliation of these texts is tagged according to the place of publication rather than to the author.

6 Translations

Nearly a quarter of the corpus texts are translations. The corpus includes translations from 69 languages, the most popular source languages being English and Russian. It is worth mentioning that sometimes Ukrainian translations are rendered from a Russian version rather from the language of the original text, and many editions do not specify this fact. When it is known that Russian served as intermediate language, this is specified in the corpus. Nevertheless, not all the translated texts within the corpus were studied with regard to a possible influence of an intermediate version.

7 Morphological Annotation

The GRAC morphology is based on the system of morphological analysis developed by the specialists of the group r2u (Andriy Rysin, Vasyl Starko and others). The system is based on the VESUM dictionary [28] and available for non-commercial use.

The program analyzes the text and for each token defines lemma and tags (grammatical markers), with ambiguity partially resolved by rule-based algorithms (see [28] for further details). An analyzed word searched in the corpus has the following format:

wordform /|lemma|/|tag1:tag2:tag3…|

The phrase *Він поспішав писати* / *Vin pospišav pysaty* ‘He wrote in a hurry’ has the following word-by-word annotation:

Він /|він|/|noun:m:v_naz:&pron:pers:3|

послішав /|поспішати|/|verb:imperf:past:m|
писати /|писати|/|verb:imperf:inf|

Thus it is possible to search by token, by lemma or by tag, and by different combinations of these.

The lemmas are marked only for the words present in the dictionary VESUM [28]. Other words can be searched only by token.

The full list of tags is available at the site of the Ukrainian Brown Corpus project [25] https://github.com/brown-uk/dict_uk/blob/master/doc/tags.txt

8 Orthography

The addition of the old texts into the corpus implies the solution of certain problems, including "correction" of the old texts in newer editions (not limited to orthography) and different orthographies in older editions. The majority of texts are included into the corpus according to modern or Soviet editions. This is shown in the metadata of the text (if known); while working with such texts one should keep in mind that they could have been altered. When it is certain that the editors did interfere, the date of the version is shown after the name of the text, while the main date of this text is still the creation date, e. g.: *Dmytro Buz'ko, Ljolja [version 2016-2018], 1924*. A minority of the texts dating back to the 19th century or to the beginning of the 20th is given in the corpus according to the older editions, the orthography being kept.

The GRAC contains texts in Skrypnykivka and Zhelykhivka, and also some texts in Yaryzhka (Russian-based orthography), such as the oldest text in the corpus (1816).

The texts in Zhelekhivka are currently only partly morphologically analyzed. The program lemmatizes correctly:

1. Orthography of the type "називати ся" (with reflexive particle written separately only in immediate postposition, cf. in the modern orthography *називатися / nazyvatysja* 'to be called')
2. Orthography of the type "цїлком" (with ї after consonants, reflecting the Western Ukrainian dialectal vocalism, cf. *цілком / cilkom* 'as a whole')
3. Orthography of the type "мякий" (without an apostrophe, cf. *м'який / m'jakyj* 'soft'):
4. Orthography of the type "сьвіт" (with a soft sign marking the regressive palatalization, cf. *світ / svit* 'world')
5. Other cases that do not correspond to the modern orthography like "моглиб" (without separated subjunctive particle, cf. *могли б* / *mohly b* '(plural) would be able'), "жити меш" (without separated futural auxiliary, cf. *житимеш žytymeš* 'you will live') and others are not recognized by GRAC.v.7, they do not have lemmas and can be found only by exact search.

9 Search Query

The GRAC search query is based on the NoSketchEngine platform [19]. The program enables search by lemma, word form and grammatical tags. Complex search queries can be built using a CQL-based query language. A user can specify text filtering (only texts of a given period, author, region, only translations from a given language etc.). It is possible to customize and save personal subcorpora with any set of textual features included into the annotation.

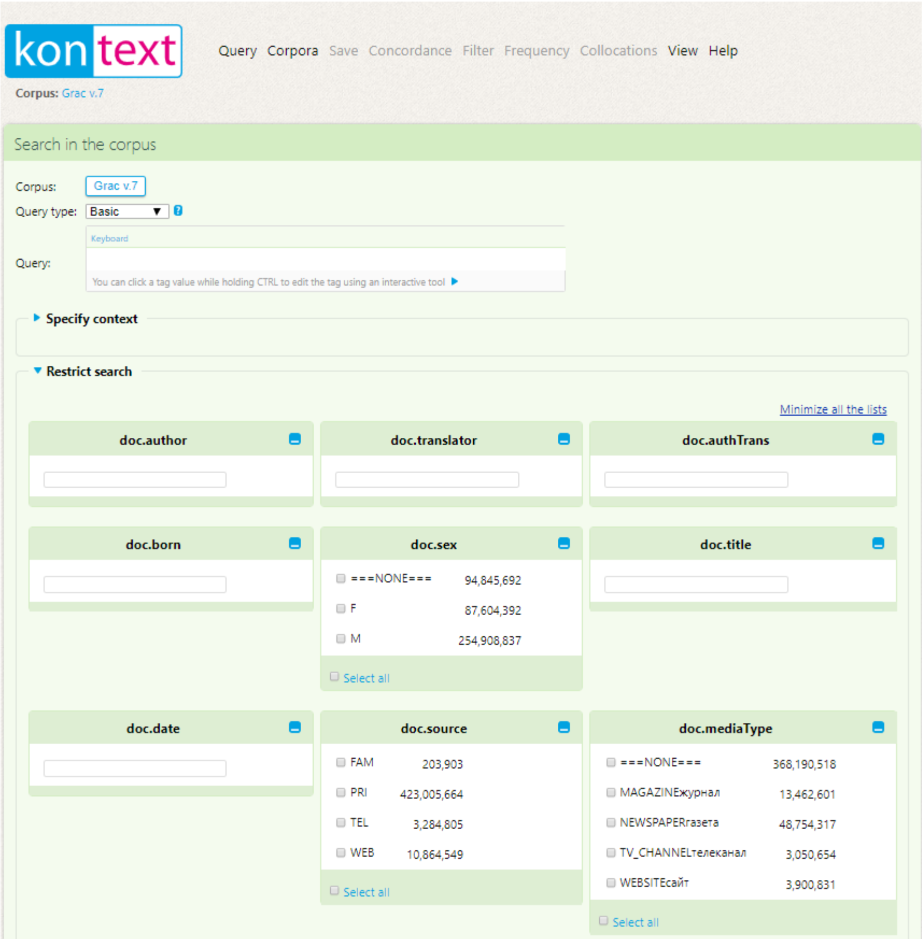
In Table 2, the metatextual attributes and the bulk of tags are presented (for the tags concerning source languages and regional markup, see the lists on the GRAC site). For the search interface see Fig. 1.

Table 1. Metatextual attributes and tags in the GRAC

Metatextual information	Attributes	Tags
Style	DOC.STYLE	FIC — fiction ACA — academic EGO — ego text JOU — journalism OFF — official SPO — spoken REL — religion FOL — folklore
Genre	DOC.GENRE	AUT — autobiography BLO — blog CHI — children DIA — diary DIC — dictionary DIS — dissertation DRA — drama EDU — education HUM — humor INT — interview LET — letter MEM — memoirs POE — poetry POP — popular PRE — speech REC — review
Academic discipline (only for ACA)	DOC.BRANCH	SOC — social sciences TEC — technical texts NAT — natural sciences

Topic (only for ACA)	DOC.THEMA	ART — art BIO — biology CHE — chemistry ECN — economics ETH — ethnography FMA — physics and mathematics GEO — geography HIS — history IT — information technologies JUR — law MED — medicine MIL — military PED — pedagogy PHL — literature and linguistics PHS — philosophy POL — political science PSY — psychology REZ — religion studies
Source language	DOC.ORIGINAL	
Date	DOC.DATE	
Publication date	DOC.PUBLICATIONYEAR	
Orthography	DOC.ORTHOGRAPHY	CONT — modern orthography ZHEL — Zhelekhivka SKRY — Skrypnykivka
Author	DOC.AUTHOR	
Translator	DOC.TRANSLATOR; DOC.AUTHTRANS for searching both original and translated texts by the same writer	
Birth date of the author	DOC.BORN	
Gender	DOC.SEX	M — male F — female
Edition	DOC.MEDIANAME	
Media type	DOC.MEDIATYPE	MAGAZINE NEWSPAPER TV_CHANNEL WEBSITE
Region	DOC.LOCCODE DOC.COUNTRY DOC.MACROREGION DOC.REGION	
Source type	DOC.SOURCE	PRI — printed source

	WEB — Internet FAM — family archive TEL — TV



Corpus: Grac v.7

Search in the corpus

Corpus: Grac v.7

Query type: Basic

Query:

You can click a tag value while holding CTRL to edit the tag using an interactive tool

doc.author

doc.translator

doc.authTrans

doc.born

doc.sex

□ ===NONE===	94,845,692
□ F	87,604,392
□ M	254,908,837

doc.title

doc.date

doc.source

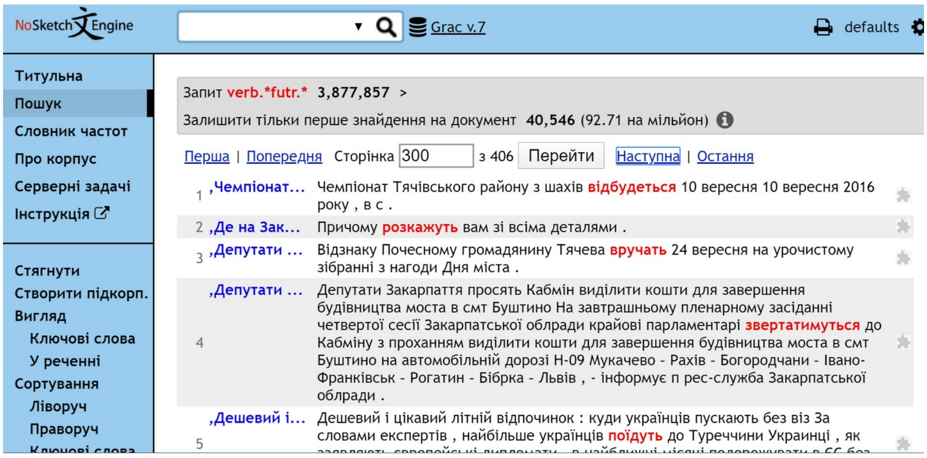
□ FAM	203,903
□ PRI	423,005,664
□ TEL	3,284,805
□ WEB	10,864,549

doc.mediaType

□ ===NONE===	368,190,518
□ MAGAZINEжурнал	13,462,601
□ NEWSPAPERгазета	48,754,317
□ TV_CHANNELтелеканал	3,050,654
□ WEBSITEсайт	3,900,831

Fig. 1. The upper part of the metatextual query form in the KonText corpus manager.

The search results are visualized as a concordance (see Fig. 2). The parameters of viewing the concordance can be additionally configured. The concordance includes contexts where the searched linguistic phenomenon is found (the contexts can be, if needed, expanded by one more sentence leftwards and rightwards by clicking on the key word) and the information on the respective sources. Users can customize the set of the information on the source that is visualized in the concordance. The full information on any text can be shown by clicking the row with metadata. The concordance can be sorted by any attribute, for example the year of creation (DOC.DATE). The concordance, configured in the way, may be downloaded as a table/database for further treatment.



Запит verb.*futr.* 3,877,857 >

Залишити тільки перше знайдення на документ 40,546 (92.71 на мільйон)

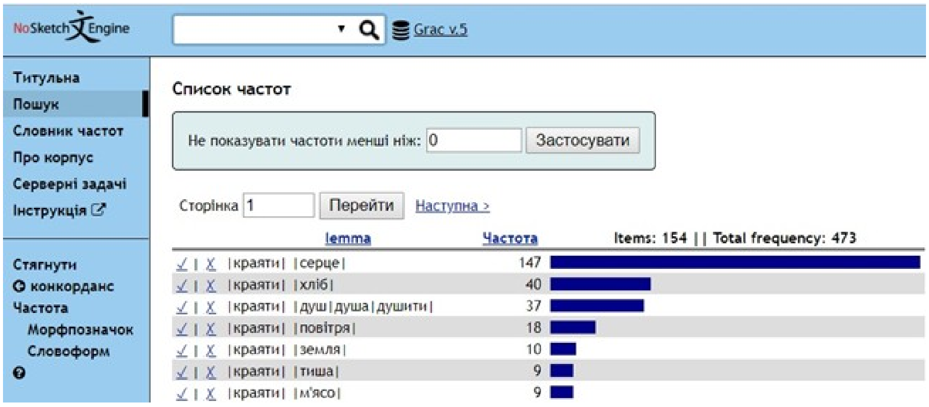
Перша | Попередня Сторінка 300 з 406 Перейти Наступна | Остання

1	,Чемпіонат…	Чемпіонат Тячівського району з шахів **відбудеться** 10 вересня 10 вересня 2016 року , в с .
2	,Де на Зак…	Причому **розкажуть** вам зі всіма деталями .
3	,Депутати …	Відзнаку Почесному громадянину Тячева **вручать** 24 вересня на урочистому зібранні з нагоди Дня міста .
4	,Депутати …	Депутати Закарпаття просять Кабмін виділити кошти для завершення будівництва моста в смт Буштино На завтрашньому пленарному засіданні четвертої сесії Закарпатської облради крайові парламентарі **звертатимуться** до Кабміну з проханням виділити кошти для завершення будівництва моста в смт Буштино на автомобільній дорозі Н-09 Мукачево - Рахів - Богородчани - Івано-Франківськ - Рогатин - Бібрка - Львів , - інформує п рес-служба Закарпатської облради .
5	,Дешевий і…	Дешевий і цікавий літній відпочинок : куди українців пускають без віз За словами експертів , найбільше українців **поїдуть** до Туреччини Украинці , як

Fig. 1. The concordance for different Future tense markers in Ukrainian.

The results of the search can be used to generate automatically frequency lists by different attributes (word form, lemma, tags, left and right context etc), frequency lists are also available for downloading. Frequency dictionaries can be generated for any subcorpus.

The example in Fig.3 shows the upper part of the frequency list for the combination of the verb *краяти / krajaty* 'cut' with different nouns in accusative ('heart', 'bread', 'soul', 'air', 'ground', 'silence', 'meat').



Список частот

Не показувати частоти менші ніж: 0 Застосувати

Сторінка 1 Перейти Наступна >

	lemma	Частота	Items: 154 || Total frequency: 473
✓ | X	|краяти| |серце|	147	
✓ | X	|краяти| |хліб|	40	
✓ | X	|краяти| |душ|душа|душити|	37	
✓ | X	|краяти| |повітря|	18	
✓ | X	|краяти| |земля|	10	
✓ | X	|краяти| |тиша|	9	
✓ | X	|краяти| |м'ясо|	9	

Fig. 1. The verb *краяти* in collocation with different nouns in accusative.

The GRAC has an additional option for building frequency plots. Several types of plots are supported (developed by Tymofij Nikolajenko).

- A frequency plot (according to any CQL query) that is build by instances per million (ipm) with regard to years. The example in Fig. 4 shows that the frequency of the variant *в Україні* / *v Ukrajini* 'in Ukraine', as it is perceived to describe a country rather than a region, increases frequency after the proclamation of the Ukrainian independence in 1991, whereas the *на Україні* / *na Ukrajini* variant decreases.

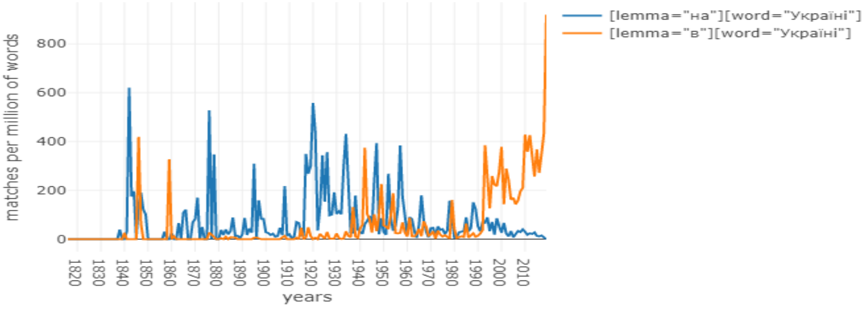


Fig. 1. The plot for the variants *на Україні* (dark color) and *в Україні* (light color) 'in Ukraine'.

The data (used to plot ipm with regard to years) for all traces are indicated as tables.

- A plot for the ratio between frequencies of more than one linguistic phenomena with regard to years. An illustration in Fig. 5 shows the distribution of the synonymous lexemes *слухавка / sluxavka* and *трубка / trubka* 'handset of a telephone'. The latter variant, as it is close to the Russian word, has been declining in frequency since 1990s.

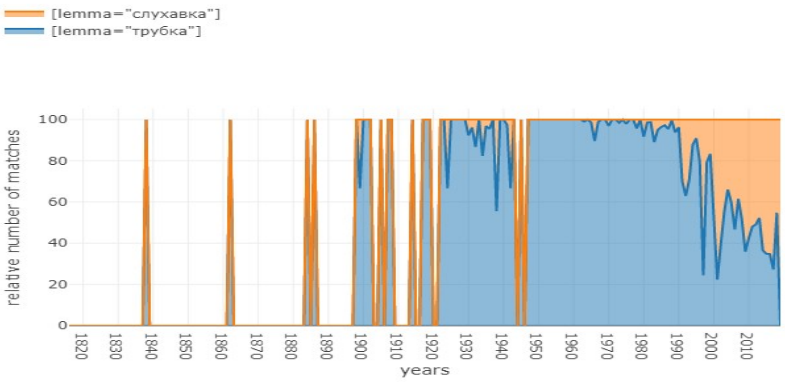


Fig. 1. Distribution of the variants *слухавка* (dark) and *трубка* (light)

All the plots may be filtered by different subcorpora, including regional ones. We may cite frequency plots showing diachronic distribution of two linguistic units in different regional subcorpora. Fig. 6 shows percentage of the variants of the preposition 'from' in the central region of Ukraine and the analogical distribution distribution for the western region. In the "central" texts, until 1930s, there is a competition between the variants *од / od* (above on both plots, brighter color) and *від / vid* (below on both plots, darker color), with the variant *від / vid* taking over as the modern standard since 1940s. In the "western" texts *од / od* is very rare throughout (and gains some momentum only in 1970s as an experiment influenced by the older "central" norm, perceived as archaic).

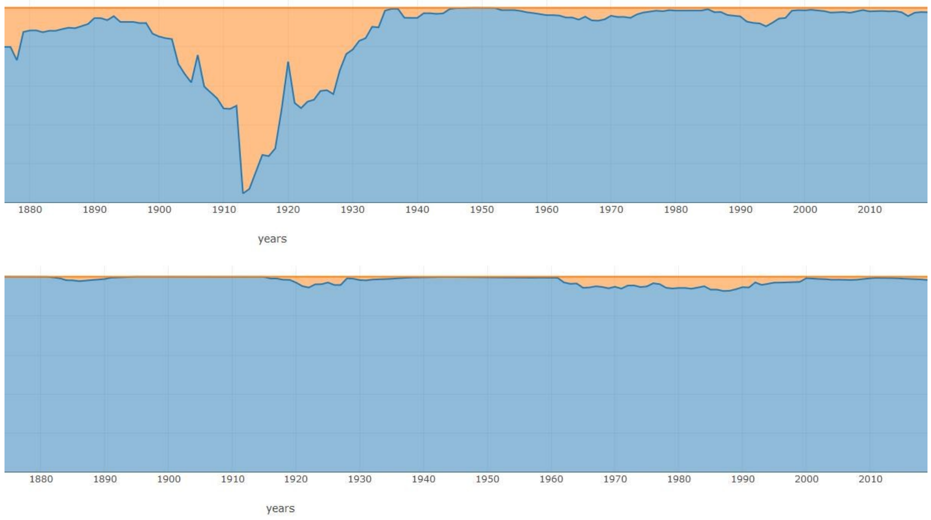


Fig. 1. *Од* and *від* in the central texts (above); in the western texts (below).

- yet another type of plots indicates the total number of words in all documents for a particular year (the so-called 'norm').

10 Conclusion

The corpus that has so detailed markup combined with modern tools for search and processing of results, is a new tool in Ukrainian studies that gives the opportunity to raise new research questions concerning the history of development of modern written Ukrainian, its regional variability, regional norms and standardization.

Future plans of its development include analyzing the current composition of the corpus and defining the representativeness gaps for eventual readjustment. As the corpus was primarily compiled using the texts easily available at the moment, it has some gaps with regard to its representativeness in such, for example the Soviet texts of 1950s-1980s, political/propagandistic and academical discourse alike. They are seldom read and digitalized, whereas they reflected many tendencies in the language (not only the imperial Russification as it is often stated) and served as sources for more modern developments. The instruments for visualization of search results are to be improved.

Special subcorpora and specific markup for them is to be further developed, including poetic corpora, parallel and perhaps spoken texts.

Directions of future evolution for GRAC include genre diversity, more detailed annotation (including semantic and possibly syntactic markup) and integration of tools for text processing.

Acknowledgements

We thank Orest Drul, Maksym Bystryckyi, Natalia Mikhailivska, Mykola Zharkykh, and everybody who creates and develops online libraries, Olena Levchenko and the Politekhnika students for prepared texts, especially for 800 letters from family archives, Vasyl Starko and the UCU students that prepared for the GRAC a collection of academic texts of 1995-2016 from the electronic library of the National Academy, Solomiia Buk and the Ivan Franko National University of Lviv students, Olena Dotsenko and the Borys Hrinchenko Kyiv University students for prepared texts, the Tempora publishing house for the interwar Ukrainian texts, Olena Yavorska and the State Literary Museum (Odesa), Mykhailo Nazarenko, Artem Fedorinchyk, Taras Shmiger, Olha Sakharova, Oksana Taran, Volodymyr Hutorov and other partners that afford texts for the corpus.

References

1. Ažnjuk, B. M.: Movna jednist′ natsiji: diaspora i Ukrajina. Ridna mova, Kyjiv (1999). [Azhniuk, B. M.: The Language Unity of the Nation: Diaspora and Ukraine. Ridna mova, Kyiv (1999)]
2. Brown corpus of Ukrainian, https://github.com/brown-uk/corpus, last accessed 2020/04/12.

3. Corpora Collection of the Leipzig University, http://corpora.informatik.uni-leipzig.de/de?corpusId=ukr_mixed_2014, last accessed 2020/04/12.
4. Corpus del Español: Web/Dialects, https://www.corpusdelespanol.org/web-dial/, last accessed 2020/04/12.
5. Corpus of Global Web-based English (GloWBE), https://www.english-corpora.org/glowbe/, last accessed 2020/04/12.
6. Corpus of Spoken Rusyn, http://parasolcorpus.org/Varchola1/login.php, last accessed 2020/04/12.
7. Corpus of the Chtyvo library, http://korpus.org.ua/, last accessed 2020/04/12.
8. Danylenko, A.: How Many Varieties of Standard Ukrainian Does One Need? Die Welt der Slaven LX, 223–247 (2015).
9. Darčuk, N. P.: Doslidnycʹkyj korpus ukrajinsʹkoji movy: osnovni zasady i perspektyvy. Visnyk KNU im. Tarasa Ševčenka. Serija: Literaturoznavstvo. Movoznavstvo. Folʹklorystyka 21, 45–49 (2010). [Darchuk, N. P.: Research corpus of the Ukrainian language: basic principles and perspectives. Bulletin of Kyiv Shevchenko University. Series: Literary Studies. Linguistics. Folklore 21, 45–49 (2010)]
10. Fokin, S.: Korpusy tekstiv: zdobutky Ukrajiny ta perspektyvy vraxuvannya zakordonnoho dosvidu. Visnyk KNU im. Tarasa Ševčenka. Serija: Literaturoznavstvo. Movoznavstvo. Folʹklorystyka 28, 51–54 (2018) [Fokin S. Corpus of texts: achievements of Ukraine and prospects of taking into account foreign experience // Bulletin of Taras Shevchenko National University of Kyiv. Series: Literary Studies. Linguistics 28, 51–54 (2018)]
11. Gritsenko P. Je. Nekotoryje zamečaniya o dialektnoj osnove ukrainskogo literaturnogo jazyka. In.: Toporov, V. N. (ed.) Philologia slavica: K 70-letiju akademika N. I. Tolstogo. S. 284–294. Nauka, Moskva (1993). [Gritsenko P. E. Some remarks on the dialect basis of the Ukrainian literary language. In.: Toporov, V. N. (ed.) Philologia slavica: Papers presented to Nikita Tolstoy on his 70th anniversary. Nauka, Moscow, pp. 284–294. (1993)]
12. International Corpus of English. http://ice-corpora.net/ice/index.htm, last accessed 2020/04/12.
13. KTUM Corpus of Ukrainian texts, http://www.mova.info/corpus.aspx, last accessed 2020/04/12.
14. Laboratorija Ukrajins’koji, https://mova.institute, last accessed 2020/04/12.
15. Mair, C.: World Englishes and Corpora. In.: Filppula, M., Klemola, J., Sharma, D. (eds.) The Oxford Handbook of World Englishes, pp. 103–122. Oxford University press, Oxford (2013).
16. Matvijas, I.: Dialektna osnova ukrajinsʹkoji literaturnoji movy. Movoznavstvo 6, 26–36 (2007). [Matviyas, I.: The dialectal basis of the Ukrainian literary language. Movoznavstvo/Linguistics. 6, 26–36 (2007)]
17. Russian-Ukrainian and Ukrainian-Russian parallel subcorpora of the RNC, http://www.ruscorpora.ru/search-para-uk.html, last accessed 2020/04/12.
18. Russkij jazyk i novyje texnologii. Novoje literaturnoje obozrenije, Moskva (2014). [Russian language and new technologies. New Literary Review, Moscow (2014).]
19. Rychlý, P.: Manatee/Bonito-A Modular Corpus Manager. In.: Sojka P., Horák A. (eds.) First Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2007, pp. 65–70. Masaryk University, Brno (2007)
20. Savčuk, S. O.: Korpus kak instrument dlja issledovanija osobennostej funkcionirovanija russkogo jazyka v regional'noj presse. Trudy Instituta russkogo jazyka RAN 6, 333–365 (2015). [Savchuk S. O. Corpus as a tool for studying the features of the functioning of the Russian language in the regional press. Proceedings of the Russian Language Institute 6, 333–365 (2015)]

21. Schmied, J.: Corpus linguistics and non-native varieties of English. World Englishes 9 (3), 255–268 (1990).
22. Švedova, M. O.: Dynamika syntaksyčnyx konstrukcij na poznačennja šljaxu pry dijeslovax ruxu: korpusne doslidžennja. Ukrajins′ka mova 3, 67–79 (2018). [Shvedova, M. O.: Dynamics of syntactic constructions for marking path with motion verbs: a corpus-based study. Ukrainian Language 3, 67–79 (2018).]
23. Švedova, M. O.: Korpusni metody doslidžennja rehional′nyx vidminnostej u mežax odnijeji movy (na materiali rehional′nyx korpusiv ukrajins′koji ta rosijs′koji mov). Visnyk Xarkivs′koho nacional′noho universytetu im. V. N. Karazina. Serija: Filolohija 77, 33–38 (2017) [Shvedova, M. O.: Corpus studies of regional differences within a language (based on the material of regional corpora of Ukrainian and Russian languages). Bulletin of V. N. Karazin Kharkiv National University. Series: Philology 77, 33–38 (2017)]
24. Švedova, M. O.: Stanovlennja prysvijnoho zajmennyka tret′oji osoby množyny v ukrajins′kij literaturnij movi. Movoznavstvo 4, 40–53 (2018). [Shvedova, M. O.: Development of the possessive third person singular in Modern Ukrainian. Movoznavstvo/Linguistics 4, 40–53 (2018)]
25. Shvedova, M., von Waldenfels, R., Yarygin, S., Kruk, M., Rysin, A., Starko, V., Woźniak, M.: GRAC: General Regionally Annotated Corpus of Ukrainian, uacorpus.org, last accessed 2020/04/12.
26. Sičinava, D. V.: Parallel’nye teksty v sostave Nacional’nogo korpusa russkogo jazyka: novye napravlenija razvitija i rezultaty. Trudy Instituta russkogo jazyka RAN 6, 194—235 (2015). [Sitchinava, D. V.: Parallel texts within the Russian National Corpus: new directions and results // Proceedings of the Russian language institute 6, 194-235 (2015)]
27. Starko, V.: Building of the Brown Ukrainian Corpus. Movni i kontseptual'ni kartyny svitu [Linguistic and Conceptual Weltbilder] 48, 415-421 (2014).
28. Starko, V., Rysin, A.: Velykyj elektronnyj slovnyk ukrajins′koji movy (VESUM) jak zasib NLP dlja ukrajins′koji movy (u druci) [Starko Vasyl (Lutsk, Ukraine), Rysin Andrew (Cary, USA). The Great Electronic Dictionary of the Ukrainian Language (VESUM) as a NLP Tool for the Ukrainian Language (forthcoming)]
29. Taranenko, O. O.: Mova ukrajins′koji zaxidnoji diaspory i sučasna movna sytuacija v Ukrajini (na zahal′noslov‘jans′komu tli). Movoznavstvo 2-3, 63-99 (2013) [Taranenko, O. O.: The Language of the Ukrainian Western Diaspora and the Current Linguistic Situation in Ukraine (against the Slavic Background). Movoznavstvo/Linguistics 2-3, 63-99 (2013).]
30. UberText, https://lang.org.ua/uk/corpora/, last accessed 2020/04/12.
31. Ukrainian national linguistic corpus, http://unlc.icybcluster.org.ua/virt_unlc/, last accessed 2020/04/12.
32. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. In.: G. Angelova et al. (eds.) Proceedings of the RANLP 2005, pp. 590-596, INCOMA, Shoumen (Bulgaria) (2005).
33. Werner, V., Seoane, E. (eds): Re-Assessing the Present Perfect: Corpus Studies and Beyond. Mouton de Gruyter, Berlin (2016).