Regional Annotation within GRAC, a Large Reference Corpus of Ukrainian: Issues and Challenges

Maria Shvedova[a], Ruprecht von Waldenfels[b]

[a] *Kyiv National Linguistic University, Ukraine*
[b] *Friedrich-Schiller-Universität Jena, Germany*

Abstract

The *General Regionally Annotated Corpus of Ukrainian* (GRAC; uacorpus.org) is a general-purpose reference corpus of Ukrainian and as such intended for a wide range of research tasks. In terms of structure, annotation and metadata it generally follows the model of existing reference corpora such as the national corpora of Czech, Russian or Polish, or the BNC. What sets GRAC apart from these corpora is the distinctive feature of regional markup. The need for such markup follows from specific properties of standard Ukrainian: due to its complex history, Ukrainian exhibits significant regional variation which has not yet been systematically investigated on the basis of a large corpus. Taking this variation into account is both a challenge for any comprehensive research into Standard Ukrainian, and constitutes an object of inquiry in its own right. In this paper, we present and motivate the principles of regional markup realized within GRAC and discuss issues of territorial representativity. We then present case studies of regional variation of Ukrainian and discuss questions and difficulties that arise in this context.

Keywords[1]

Ukrainian language, corpus, regional variation.

1. Introduction

A linguistic corpus designed for a certain research question contains specifically collected data in sufficient quantity. A reference corpus [21] such as GRAC, in contrast, is intended to be a universal tool for a wide range of research questions (other examples for reference corpora include the British, Czech, Russian or Polish national corpora, the Slovene Gigafida or the German DWDS and DeReKo corpora). The issue of contents, annotation and representative balancing of text types in such a large reference corpus is largely a general theoretical question which is independent of specific research tasks and potentially even independent of the language of the corpus. However, the practice of building GRAC shows that properties of the language in question can make certain modifications of this universal structure necessary.

In this article we discuss the problem of design and balancing of a large reference corpus of Ukrainian. We describe the regional structure of GRAC, conditions of its creation and possibilities of its use.

A reference corpus (often called national corpus) is intended to be a universal tool for a wide range of research tasks. Thus, the problem of representativeness and balance of the corpus is of high relevance, and corpus linguists often refer to it [19, 16, 20, etc.]. The bulk of the corpus usually consists of fictional, journalistic and academic texts in various proportions that are, more or less conventionally, designed to provide a corpus snapshot that is representative of the standard language in question, even though its contents may not strictly correspond to the proportions of different types of texts that are actually created within the language community.

[1] COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: corpus.textiv@gmail.com (M. Shvedova); ruprecht.waldenfels@gmail.com (R. von Waldenfels)
ORCID: 0000-0002-0759-1689 (M. Shvedova); 0000-0001-5822-5040 (R. von Waldenfels)

In 1993 Douglas Biber wrote that a corpus that would be consistent with the real language practice in a community would have "roughly 90% conversation and 3% letters and notes, with the remaining 7% divided among registers such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writing. (Very few people ever produce published written texts, or unpublished written and spoken texts for a large audience.)" [19, p. 247]. However, it is clear that by the 2020s, with internet and electronic publications widely available, this proportion has changed: people produce far more written texts, including those accessible to a large audience.

The statistical representativeness of the corpus of traditional written texts is also problematic, as contemporary researchers have pointed out. It is practically impossible to assess the correspondence of a corpus to the linguistic reality, since there are no mechanisms that would allow us to accurately measure the representation of different texts. Some parts of the corpus may be considered representative (for example, a sufficiently large collection of contemporary newspapers), but they would be representative of their period, style, and genre, and not of the language as a whole [20]. Practically, the most reliable corpus in terms of representativeness is one that enables us to work with a maximally wide range of text types, and with as large and diverse samples of these type as possible.

In Ukrainian linguistics, attempts have been made to transfer the structure adopted in various reference corpora to a Ukrainian language corpus. The "Ukrainian Language Corpus" project was conceived by the team of the Computer linguistics laboratory at the Institute of Philology of the Kyiv National Taras Shevchenko University under the supervision of Natalia Darchuk [11, 4]. It contains large subcorpora of fiction, journalism, and academic texts, not unlike the reference corpora of other languages. In a number of theoretical works, Orysia Demska proposed a complex structure of a future Ukrainian national corpus which takes into account a large number of variables ranging from standard attributes such as author, time and genre to rather specific attributes such as conditions of a conversation, level of preparedness, intended audience, or education, profession and place of work of the author [6]. However, this rather complex structure has so far not been realized in an actual large corpus.

The practical experience of corpus building and the analysis of a large number of texts during this process shows that, in fact, a wide range of factors influences linguistic phenomena on the level of individual texts. This includes external characteristics concerning the author, such as their level of education, their profession and their political views; more difficult to capture characteristics such as aesthetic preferences, style, and identity projection; and many more. Besides, textual factors such as genre, topic, register, stylistic markedness, and others play a role in addition to diachronic, geographic and other factors. Generally, the range of possible factors constitutes an open set the members of which cannot all be envisaged in the metatextual information or represented by a sufficient number of texts in the corpus. The list of metatextual attributes is thus bound to be insufficient.

On the other hand, the available metatextual information that is captured by such attributes will necessarily sometimes be inaccurate. For example, the date of texts may be inaccurate, since newer texts often contain quotations from older texts. The meta data concerning genre may also be misleading: mass media largely contain journalistic texts, but may also publish short fiction, letters to the editor, poems, speeches, official documents, the tv program, the weather forecast and so on. All of these texts would theoretically need to be extracted and processed separately in order for the meta data to be completely accurate. It is very difficult to trace and differentiate all such cases in a large corpus – a Herculean task that may slow down data collection to a crawl.

Some other theoretical ideas expressed in the literature about the composition of a national corpus of Ukrainian language also seem to be rather difficult to implement in practice. For example, O. Demska [5] proposes to include dialectal material in such a corpus and to exclude Surzhyk, a spoken variant of Ukrainian that exhibits strong effects of interference with Russian. Such an approach may be difficult to implement in practice and raises a number of difficult to solve issues. For example, in what kind of transcription should the dialectal material be introduced? A normalized transcription may remove exactly those phonetic and dialectic features that makes these variants interesting in the first place. A phonetic transcription, on the other hand, creates an additional challenge for the automatic grammatical annotation, which in a standard-oriented corpus cannot handle dialectal features. The inclusion of such texts is thus a task for specialized corpora. On the other hand, however, a standard-oriented corpus cannot completely exclude dialectal traits or other elements of vernaculars and mixed speech like Surzhyk. These elements appear in texts that more or less directly reflect spoken language, ranging from fiction to parliamentary transcripts. How can one accurately determine the extent of mixed Russian elements in the text, so that it could be qualified as clearly Surzhyk and

excluded from the corpus? Most often we encounter Surzhyk in the corpus in the form of isolated elements: in the speech of characters, linguistic games or puns, examples in linguistic publications, and so on. This is a stylized Surzhyk that forms the integral part of a literary or academic text. It is clear that a standard-oriented language corpus cannot be a full-fledged tool for the study of other linguistic varieties, a purpose for which special oral corpora with sound recordings are designed. But it is impossible to avoid the elements of oral speech in a large corpus, and such an objective is utopic in itself. At the same time, we fully agree with O. Demska when she points out the necessity of creating a separate corpus of Ukrainian Diaspora language and elaborates prospects of researching this variant on the basis of the corpus [6].

We believe that when creating a reference corpus for a particular language, we must not only focus on universal aspects of the structure of a reference or national corpus. Such universal aspects concern some sort of representative sample of what are deemed the main genres and text types relevant for a language or language variant. Note that some text types rarely find representation: for example, user manuals, legal or official documents such as passports or invitations to Parent-Teacher-Meetings are included by far not in all reference corpora – even though the language of such texts is clearly part of a speaker's linguistic knowledge. Note that GRAC contains some of such poorly represented text types, for example, Kyiv trolleybus ticket from the 1960ies; however, but for the moment, such text types are not focused on, as they are generally rather difficult to collect and other texts take precedence. In general, these are issues that are relevant in any reference corpus of a major standard language.

Aside from such universal issues, it is also important to take into account the specific features governing the variability of a particular language when designing a reference corpus for this language, as they should be available for study in the corpus. In the case of Standard Ukrainian, a factor of prime importance is geography: Ukrainian was a polycentric language for a long period, an important factor behind the emergence of some modern lexical and grammatical variants.

The design of GRAC thus reflects universal aspects of a reference corpus in its textual composition, its typology of text types and in the range of metadata that is provided. In addition to this, it contains regional annotation, which is important for Ukrainian and reflects its polycentric status in past and present (as concerns diaspora Ukrainian) as well as its complicated history of standardization.

2. Basis and principles of regional markup within GRAC

The Modern Ukrainian standard language (often called literary language in the Slavic tradition) finds its beginnings at the end of the 18th century. It was originally based on Southeastern, Southwestern and Northern dialects that had significant differences to each other. The political division of the Ukrainian people between the Russian and the Austro-Hungarian empire before WWI and the influence of different *dachsprachen* (roofing languages – mainly Polish in the West, Russian in the East) in distinct cultural centers deepened these linguistic differences. For the 19th to early 20th century, Hrytsenko [10] thus distinguishes the variants of Middle Dnieper, Galician, Bukovinian, Transcarpathian, Ruthenian. In the language of classical Ukrainian literature of the 19 and early 20th centuries one can find clear features of the writers' native dialects, especially dialectal vocabulary, phonetics, grammatical forms to varying degrees – Western Ukrainian authors are more noticeable in this respect. The Western Ukrainian texts generally reflected a more archaic dialectal syntax of southwestern dialects to some extent [13]. In the late nineteenth – early twentieth century, the mutual influence of the two major literary standards, Middle Dnieper and Galician, intensified, finally leading to what one can call territorial variants of a single standard language [10]. A comprehensive common spelling and grammatical standard of the Ukrainian language was first codified only in 1928; this standard was later modified in the 1930ies in a top-down attempt to move it closer to the Middle Dnieper variant and Russian, a development not shared by writers of the influential Ukrainian diaspora in the West after WWII. Independence after 1991 again lead to further changes, partly rolling back those of the 1930ies. Because of this complex standardization history, Ukrainian still exhibits significant lexical and grammatical variability today, arguably more than many other standard languages.

An important aim of GRAC is to provide an instrument that would enable us to trace these linguistic regional differences in the historical part of the corpus on a sound empirical basis and see whether and to what extent they are preserved in modern texts today.

The regional markup of the corpus is based on the contemporary administrative structure of Ukraine. This is partly because of pragmatic reasons: administrative borders are clearly defined and it is possible to look them up in standard sources. While the administrative structure does not necessarily reflect the dialectal landscape of Ukraine, this choice does have a sociolinguistic dimension since the administrative regions do present socioeconomic and cultural entities of some relevance that are typically oriented towards the same centers. These administrative regions are then united in macroregions consisting of the Western, Eastern, Central, Southern and Northern area. Kyiv as the capital with people coming from different regions is treated as a separate macroregion. Below are the graphs showing how our texts are distributed across these macroregions overall in the corpus (Fig. 1) and across time (Fig. 2). Kyiv and the Western macoregion are represented by the largest numbers of texts. The other regions have much less texts.

Macroregion	Tokens	%
W	172303252	46
KYV	118565515	32
E	26624696	7
C	23900708	6
S	16903552	5
N	12944789	3

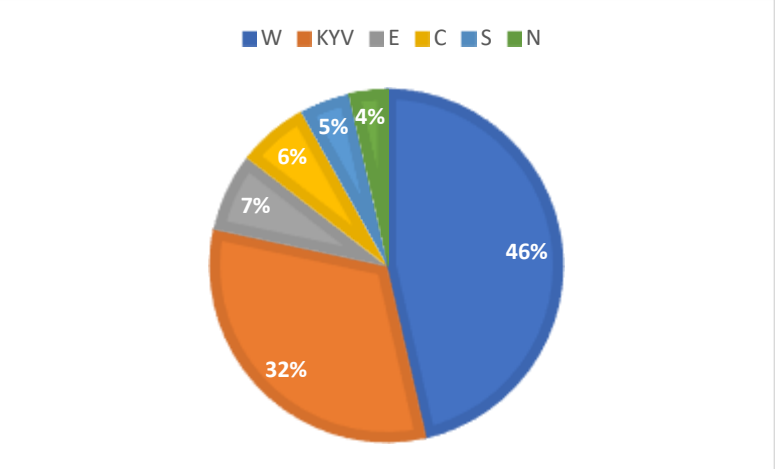

Figure 1: Composition of GRAC by macroregions

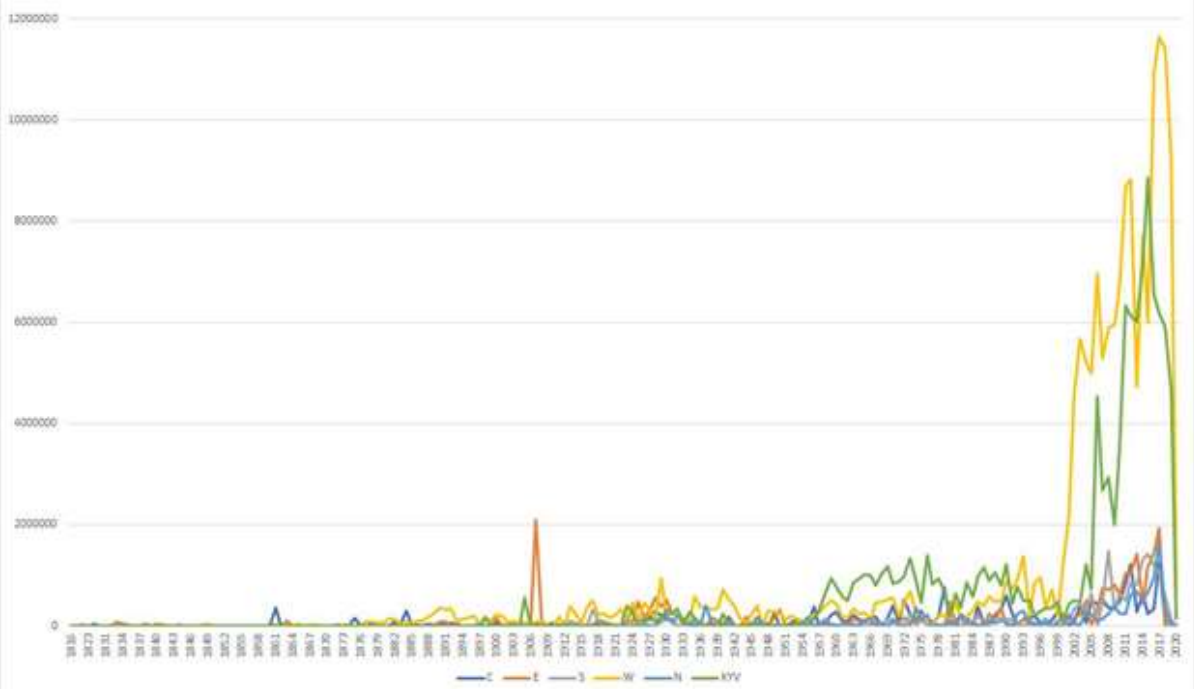

Figure 2: Distribution of tokens by macroregions and years

Media texts (papers, news sites on the web) are marked by the region where the respective media appeared. Other texts are annotated by the region where the author (or the translator, for a translated text) was born, studied or lived for more than ten years.

The regional annotation is thus generally linked to the author of a text where such an author is available. 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. In the process of annotation, biographical information from all kinds of sources is evaluated so that the regional annotation reflects the Ukrainian linguistic biography of the author as closely as possible. For example, the writer Emma Andijewska was born in Donetsk (Stalino at the time), as a child moved to Kyiv Region and then emigrated to Western Europe. Accordingly, she was first assigned three places: Donetsk, Kyiv region and Germany. However, since she herself stated that she first came into contact with Ukrainian in Kyiv Region, Donetsk was subsequently dropped. Since the regional annotation of texts is linked to their authors, all of Andijewska's texts now have Kyiv Region as first, and Germany as second region – regardless of where they were actually written. A more fine-grained annotation seems hardly feasible.

Approximately 85.5% of GRAC v.10 is annotated by region. Texts created in Ukraine that have one macroregion make up 60% of GRAC v.10 corpus.

For regional text markup, GRAC has the attributes DOC.COUNTRY, DOC.MACROREGION (North, West, South, East, Center, Kyiv: Fig. 3), DOC.REGION, and DOC.LOCCODE, which for convenience contains a set of all regional attributes (for example, DOC.COUNTRY = "UA", DOC.MACROREGION = "C", DOC.REGION = "CRK", and DOC.LOCCODE = "UA-C-CRK").

Figure 3: Macroregions of Ukraine in GRAC

DOC.LOCCODE for Ukraine:

UA-C-CRK — Cherkasy oblast
UA-C-KRV — Kirovohrad oblast
UA-C-KVS — Kyiv oblast
UA-C-PLT — Poltava oblast
UA-E-HRK — Kharkiv oblast
UA-E-SUM — Sumy oblast
UA-KYV-KYV — Kyiv
UA-N-CRG — Chernihiv oblast
UA-N-RVN — Rivne oblast
UA-N-VLN — Volyn oblast
UA-N-ZHT — Zhytomyr oblast
UA-S-DNC — Donetsk oblast.
UA-S-DNP — Dnipropetrovsk oblast
UA-S-HRS — Kherson oblast
UA-S-KRM — Crimea
UA-S-LGN — Luhansk oblast
UA-S-MKL — Mykolaiv oblast
UA-S-ODE — Odesa oblast
UA-S-ZPR — Zaporizhia oblast
UA-W-CRV — Chernivtsi oblast
UA-W-HML — Khmelnytskyi oblast
UA-W-IFR — Ivano-Frankivsk oblast
UA-W-LVV — Lviv oblast
UA-W-TRN — Ternopil oblast
UA-W-VNC — Vinnytsia oblast
UA-W-ZKR — Zakarpattia oblast

Aside from the above macroregions, the countries of the Ukrainian diaspora (the United States, Canada, Poland, Germany, the UK, France etc.) are distinguished in the annotation. DOC.LOCCODE for the Ukrainian diaspora starts with D, followed by a code for post-Soviet countries (DOC.MACROREGION = "V") and other countries (DOC.MACROREGION = "Z"). The third code specifies the country. For the neighboring Russia, Poland and Czechoslovakia, a fourth code is available to specify further details.

D-V-BY — Belarus
D-V-GE — Georgia (country)
D-V-KZ — Kazakhstan
D-V-MLD — Moldova
D-V-RU — Russia
D-V-RU-KBN — Kuban
D-V-RU-SSL — Eastern Slobozhanshchyna
D-V-TKM — Turkmenistan
D-Z-AR — Argentina
D-Z-AT — Austria
D-Z-AU — Australia
D-Z-BE — Belgium
D-Z-BR — Brazil
D-Z-CA — Canada
D-Z-CH — Switzerland
D-Z-CZE — Czech Republic
D-Z-CZE-SVK — Czechoslovakia (before 1992)
D-Z-DE — Germany
D-Z-EET — Estonia
D-Z-ES — Spain
D-Z-FR — France
D-Z-GB — United Kingdom
D-Z-IL — Israel
D-Z-IT — Italy
D-Z-LT — Lithuania
D-Z-LV — Latvia
D-Z-PL — Poland
D-Z-PL-HLM — Kholm region
D-Z-RO — Romania
D-Z-SRB — Serbia
D-Z-SVK — Slovakia
D-Z-SWE — Sweden
D-Z-USA — United States

3. Variability in the corpus: different factors

The regional annotation in GRAC allows us to explore different distributional patterns of variants. Here, we illustrate three different factors behind variability of Modern Ukrainian. First, the fate of erstwhile dialectal variants and how they are represented in Standard written Ukrainian today. Then, we look at the historical variants of Standard Ukrainian that, despite their merger since 1920s, still have weaker repercussions on multiple linguistic parameters. Finally, we discuss the language of Ukrainian diaspora as a phenomenon keeping many pre-WWII linguistic features and at the same time treating borrowings differently.

3.1. Variability in the corpus: the influence of dialects

The regional divisions annotated in GRAC is based on administrative boundaries and as such do not fully correspond to the dialectal map of Ukrainian. This is the case because GRAC is a corpus of texts oriented to the literary standard. As such, dialectal authenticity cannot be expected from such modern written texts of different regions. But there is the possibility to find indirect traces of dialectal influence on the regional written language. These traces form the basis of regional language variability. Such regional variability is characteristic of many European languages [18], and the regional marking of GRAC allows us to explore the Ukrainian language in this respect. In this chapter, we adduce four case studies that illustrate such an approach and show very different patterns of variability.

3.1.1. The distribution of words for 'potato' in dialects and in GRAC

The dialect atlas gives different variants for 'potato' in Ukrainian dialects: *картопля* (borrowed via Russian and Polish from German, Russian *картофель*, Polish *kartofel*, German *Kartoffel* [7]), *бараболя* (from Czech *brambor* – both phonetic variations of the name of the German province of Brandenburg, through which potatoes spread to the east [7]), *бульба* (apparently borrowed from the Polish language, and there, from Latin [7], cf. also Belarusian *бульба*), *pina* (Proto-Slavic *rěpa*, historically the name for 'turnip' [7]), *мандибурка* (from the name of Magdeburg or, according to

F. Miklosich, also Brandenburg [7]), *крумплі* (Ukrainian *кромпель* – borrowing through Polish mediation from the Slovak language; Slovak *krompl'a, krumpl'a, krumpel* [7], according to another version, in the Transcarpathian Ukrainian dialects the form of *krumpli* indicates the influence of the Hungarian language [3]) etc.

Three variants are widespread throughout Ukraine: *картопля, бараболя, бульба.* Podillya is a region of predominant use of the *бараболя* variant. In Polissya, next to the *картопля* variant, the Atlas shows a significant spread of the *бульба* variant. The greatest variety of potato names is recorded in the West: in addition to the names common in other regions, *картопля, бараболя, бульба*, in Galicia and Bukovina *мандибурка, гарбуз, біб* are attested, in Transcarpathia – *крумплі* [1].

In contrast to the dialect speech reflected in the Atlas of the Ukrainian language, in the written texts presented in GRAC one can see one main variant of *картопля*, which predominates in all the macroregions (Table 1). Of all the cases where different names for 'potato' are attested, *картопля* accounts for 90% of the use in the texts of Kyiv, East and Center, 80% in the texts of the South and West, but only 58% in the texts of the North, where the main lexeme *картопля* has a strong competitor, *бульба*. The variants *бульба* and *бараболя* are available in the texts of each macroregion. In addition, in the texts of the West there are variants of *мандибурка* (22 times) and *крумплі* (40 times). We counted only texts created in Ukraine that are tagged by a single macroregion (amounting to 60% of GRAC v.10 corpus version).

Table 1

The number of finds of different names for 'potato' in GRAC by macroregions

	N	KYV	E	W	S	C
картопля	858	3867	759	7346	451	911
бараболя	40	53	25	474	39	53
бульба	453	343	43	1307	74	51
мандибурка				22		
крумплі				40		

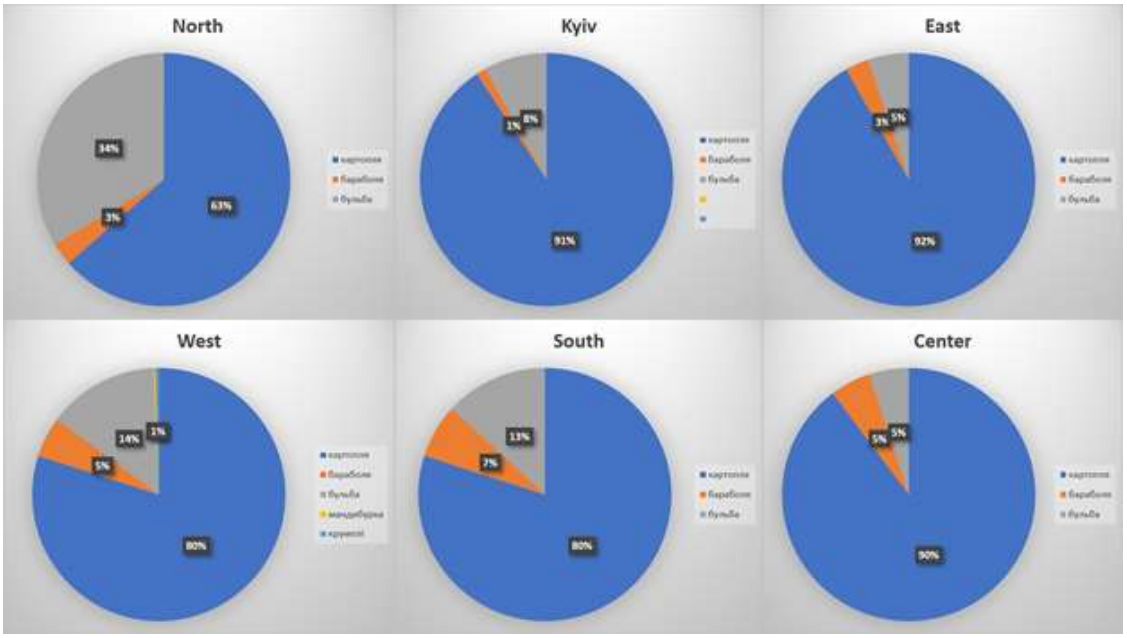

Figure 4: Distribution of words for 'potato' in GRAC by macroregions: blue for *картопля*, orange for *бараболя,* grey for *бульба,* yellow for *мандибурка,* light blue for *крумплі*

Comparing these results with the dialect map, we see that a) in written texts the standard version of *картопля* is more common than in oral speech, b) the influence of local regional oral speech on written texts is noticeable: the fact that in the texts of the North a significant share of the variant *бульба* is attested (cf. the Atlas of the Ukrainian language, as well as Belarusian *бульба*), and in the

Western texts *мандибурка* and *крумплі* are found, corresponds to the dialectal distribution of these words.

Some cases of the use of the words *мандибурка* and *крумплі* have been found outside the Western region proper, in the texts of the authors who moved outside the territory (that is the texts with several regional tags, one of which being Western), as well as those that have nothing to do with the West. About 15% of the corpus texts have more than one regional tag (showing that the authors lived for a long time in different regions or countries). Below, we consider these cases, as well as the cases where the Western Ukrainian words are attested in the texts by non-Western Ukrainian authors, in more detail.

***Мандибурка* 'potato'**. The corpus gives 30 examples, mostly from texts belonging to the Western macroregion:

*У великих горшках, узятих із двірської челядної кухні, вариться «**мандибурка**», в інших кипить окріп на стиранку, пряжиться молоко, в ринці смажиться сир (Іван Франко, Гриць і панич, 1898, UA-W-LVV). Казьо мав учителів з міста, яких щодня привозили панською повозкою; мене підготовляв місцевий учитель за пів кірця кукурудзи і корець **мандибурки**, які йому офірував мій батько (Андрій Чайківський, Прокляття, 1929, UA-W-LVV). Юрій обернувся так, аби тато бачив рух його губів, і сказав: — **Мандибурки** ще є трохи (Михайло Івасюк, Серце не камінь, 1978, UA-W-CRV). Восени капуста є на полі, буряки, морква, **мандибурка** — тобто картопля (Роман Федорів, Єрусалим на горах, 1992, UA-W-IFR & UA-W-LVV). – Що їстимеш? Може, вареників зварити чи юшку курячу, чи, може, колочену **мандибурку** або кулешу. (Ярема Ткачук, Буревії. Книга памяті, 2004, UA-W-IFR & UA-W-LVV).*

In some cases, the word *мандибурка* occurred in texts marked by several regions, one of which being the Western one.

*— Не хвилюйся, батя, зараз будемо пекти **мандибурку** і смажити кабаки (Тимофій Гаврилів, Вийди і візьми, 2014, UA-W-IFR & UA-KYV-KYV).*

Two authors use this word in lists of other names for 'potato'

*Картопля, **мандибурка**, в лушпиннях парує саме, чекаєш, поки відпарує, щоб її із миски вихопити і можна було, щоб рук уже не впекти (Андрій Кондратюк, Краса зникаюча і вічна. Т. 1, 2007, UA-N-RVN & UA-W-LVV). Печуться крумплі (картопля, **мандибурка**, бараболя, бульба, як ще?) (Ніна Бічуя, Великі королівські лови (збірка), 2011, UA-KYV-KYV & UA-W-LVV).*

The variant *мандибурка* was once found in the text by a Soviet novelist Mykhailo Stelmakh, although, according to the Atlas of the Ukrainian Language, this variant is not widespread in Podillia, where the author came from and where his novels are set. The variant in found in the speech of Maria, a village woman:

*— Як прийдеться за чужою пряжею пучки протирати, **мандебуркою** давитися, за сніп жати, тоді не раз матір згадаєш. А за Дмитром будеш жити господинею! Господинею, а не наймичкою, не поденщицею! (Михайло Стельмах, Велика рідня, 1951, UA-W-VNC & UA-KYV-KYV).*

The main name for 'potato' in Stelmakh's novel "Great Family" is the word *картопля* (19 times), *бульба* occurs only as the name of a Belarusian dance, *мандибурка* is used once, possibly as a historical marker.

***Крумплі* 'potato'.** This variant, according to the Atlas of the Ukrainian language, is widespread in Transcarpathia, and the corpus gives the word *крумплі* and its derivatives mainly in Transcarpathian texts:

*На вечерю тогди має бути 7 або 9 потрав, як: пасуля, ленча, горох, печениці, гриби, біб, **крумплї** і паленята із оліїйом (Юрій Жаткович, Замітки етнографічні з Угорської Руси, 1896, UA-W-ZKR). — Треба мені п'ять центнерів пшениці, п'ять центнерів тенгериці, хоч два вози **крумплів**, щоб міг перезимувати з дітьми (Петро Лінтур, Зачаровані казкою: Українські народні казки Закарпаття, 1969, UA-W-ZKR). – Богонько нам подарував красну днину – йдемо у Мочар **крумплики** обгортати (Дмитро Кешеля, Осінь Великих Небес, або Прирічанські характери, 2005, UA-W-ZKR). Доки відчинялися шинки, я встигав наловити плетінку раків, яку вимінював на жменю кукурудзяної муки чи пару-другу **крумплин** (Мирослав Дочинець, Криничар. Діяріюш найбагатшого чоловіка Мукачівської домінії, 2012, UA-W-ZKR).*

A smaller part are examples from texts describing Transcarpathia by authors unrelated to the Transcarpathian region. This is a metalinguistic use of the word as a description of the local dialect:

*Перевал краси. І вже рідна моя Трансільванія. Гори розступились. Розлогі долини налиті сонячною синьою млою. Трансільванія картоплі копає. **Крумплі** (Олесь Гончар, Щоденники,*

1967). Коли свого часу Інна Кваковська вийшла заміж за закарпатського хлопця, чоловік вчив її особливостей мови: «Ти що, не знаєш, що гарбузи, – дивувався він, – то по-нашому ***крумплі****?» ("Високий замок", 2012, UA-W-LVV).*

This word was once found directly in a linguistic text describing local lexicon:

У літературній мові не належать до стійкої лексики також слова, зрідка використовувані з стилістичною метою, що не набули загальнонаціонального значення і вживаються лише в місцевих говорах, як наприклад, газда (господар), когут (півень), кияхи, пшеничка (кукурудза), ярець (ячмінь), ріпа, ***крумплі*** *(картопля), верета (рядно), болоння (толока), борзо (швидко), хоч у місцевих говорах вони й виявляють велику стійкість (Михайло Жовтобрюх & Борис Кулик, Курс сучасної української літературної мови. Ч. 1, 1965).*

It is clear that for such cases the region of creation of the text is irrelevant.

Thus, it is seen that the use of regionally-marked words is relatively higher in the written texts of those regions where they are common in oral speech. But this does not mean that they cannot occur in the texts from other regions. Authors may well use them to describe the local language or borrow them for some other specific purpose. Obviously, we must also take into account the frequency.

3.1.2. The distribution of variants of a single preposition in dialects and in GRAC

Consider another example: the use of prepositions *від/од* by region. *Від/од* is historically a single preposition (Old East Slavic *отъ*) which had different variants in different Ukrainian dialects due to historical phonetic transformations. Now the main standard variant in the Ukrainian language is *від*, historically characteristic for the Southwestern dialects, and a less common variant *од*, which is used in other Ukrainian territories and is the main variant in Polissya dialects [1].

In GRAC v.10 we find 1,759,355 uses of *від* and 85,598 cases of *од*. In Western Ukrainian written texts (DOC.MACROREGION="W"), according to the oral use in the regions in question, *від* is practically the only option:

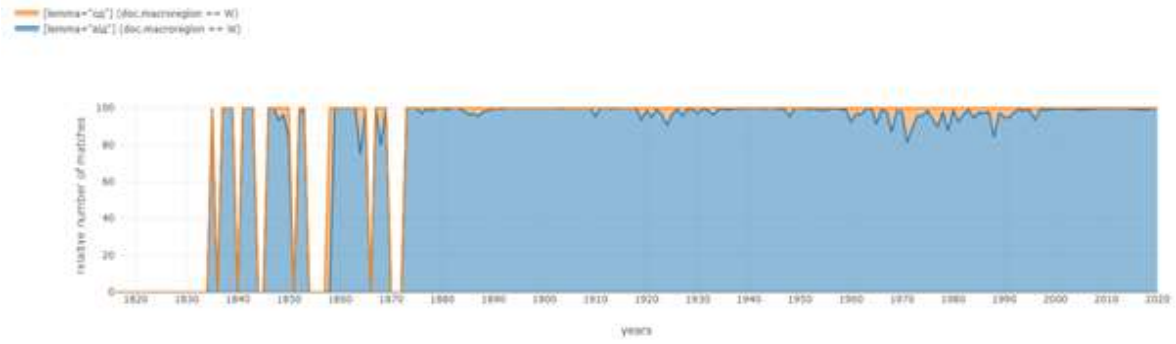

Figure 5: Prepositions *від* (blue) and *од* (orange) in GRAC, macroregion West

In the texts from other macroregions, the share of the *од* variant is considerably greater, especially before the Soviet standardization of the 1930s, which approved the *від* variant as the main norm.

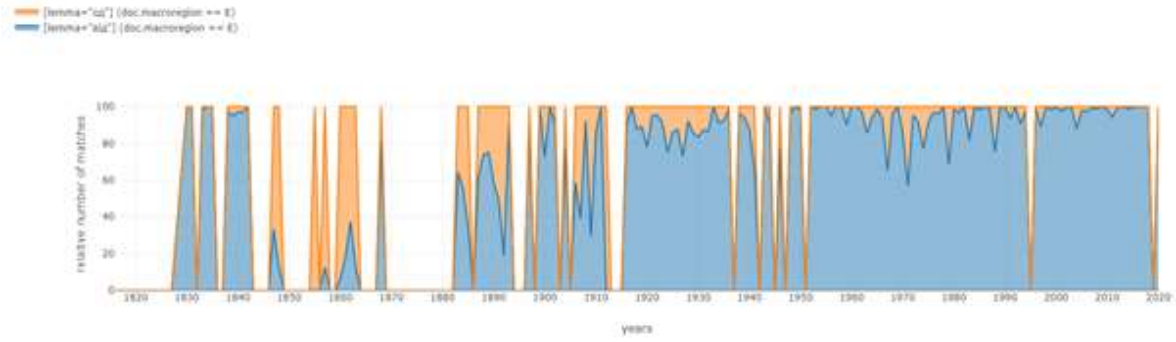

Figure 6: Prepositions *від* (blue) and *од* (orange) in GRAC, macroregion East.

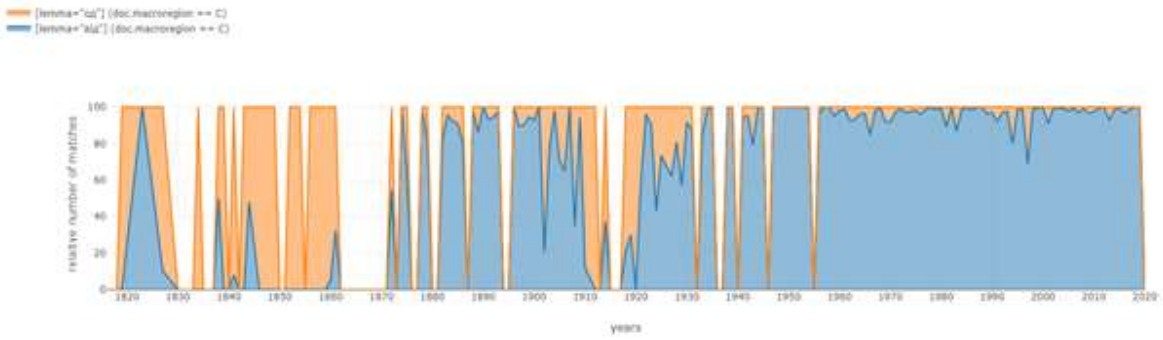

Figure 7: Prepositions *вiд* (blue) and *од* (orange) in GRAC, macroregion Center

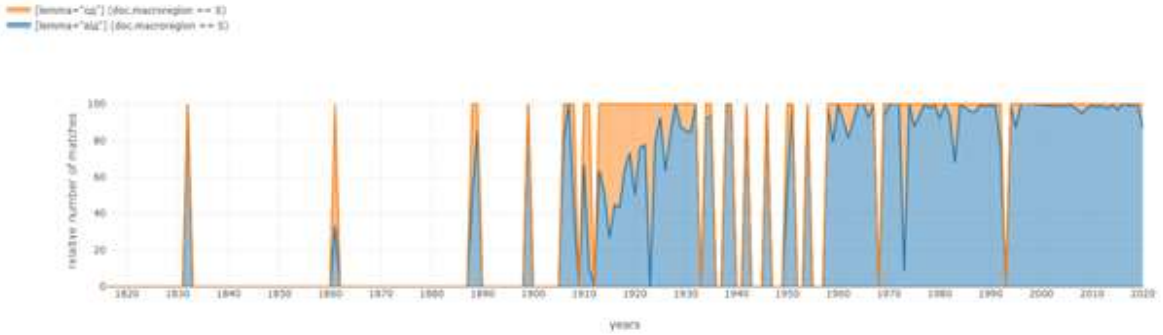

Figure 8: Prepositions *вiд* (blue) and *од* (orange) in GRAC, macroregion South

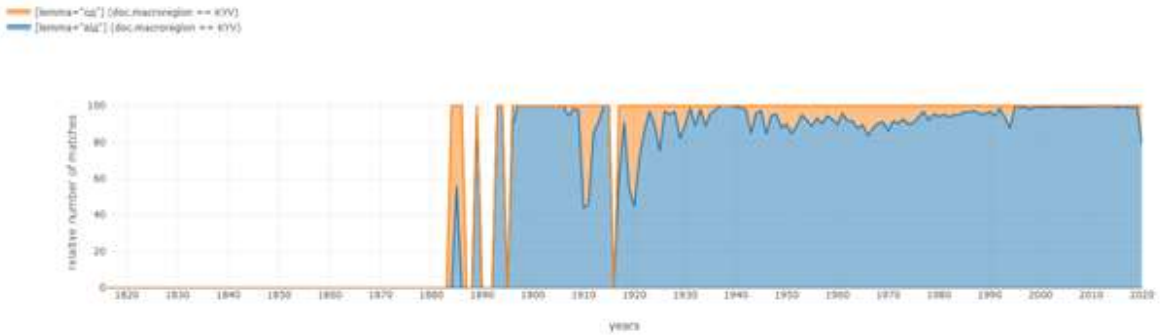

Figure 9: Prepositions *вiд* (blue) and *од* (orange) in GRAC, macroregion Kyiv

Unfortunately, the texts of the Northern region, where the Polissya dialects are characterized by the *од* variant, are in GRAC v.10 less than all others.

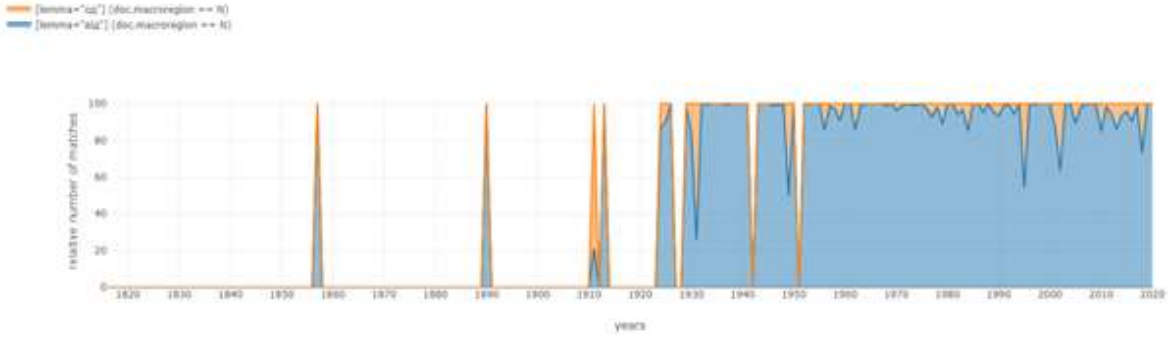

Figure 10: Prepositions *вiд* (blue) and *од* (orange) in GRAC, microregion North

3.2. Influence of old variants of the standard language on variability in the corpus

Many phenomena are practically unaffected by the regional parameter.

For example, attributive noun phrases with different word order: “adjective + noun” (as in *добрий день*) vs. “noun + adjective” (as in *день добрий*) (Fig. 11, 12).

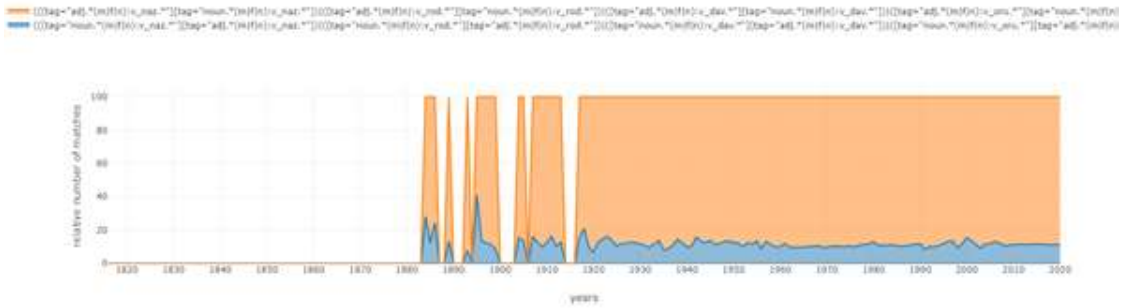

Figure 11: Phrases with the word order: "adjective + noun" (orange) vs. "noun + adjective" (blue) in Kyiv

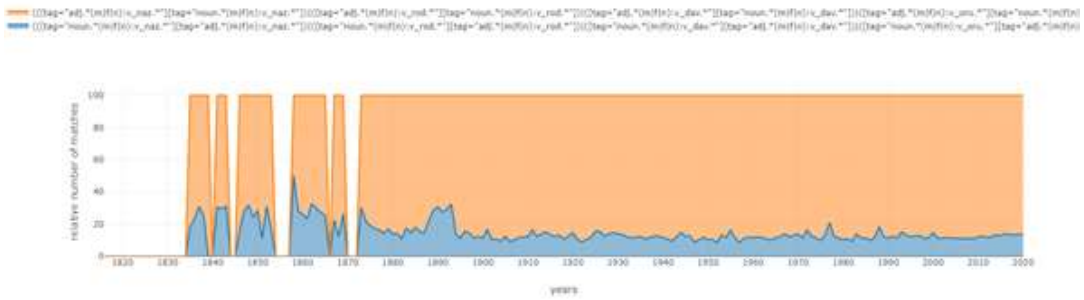

Figure 12: Phrases with the word order: "adjective + noun" (orange) vs. "noun + adjective" (blue) in the West

However, some differences in such graphs can be seen for the texts of the 19th century, when the influence of the Polish model (noun+adjective) in the Western Ukrainian variant of the standard language was noticeable:

— Що нам там займатися науковими питаннями, філософією, економією, ***науками природничими****! (Іван Франко, Наша публіка, 1888, UA-W-LVV). Що се ви говорите про якісь загальноукраїнські справи, про конечність вироблення одностайної галицькоукраїнської* ***мови літературної****! (Іван Франко, Наша публіка, 1888, UA-W-LVV). Видко, що* ***елемент польскій*** *на Шлеску о много слабшій, чим нїмецкій, а навіть моравскій. Він розкладаєсь під впливом тих двох* ***культур національних*** *і колись загине цілком тим певнїйше, що про селянина-Поляка там не дбає нї священик, нї учитель (Як колишний польскій Шлеск нїмчиться. — Міщаньска "Руска Бесїда" в Снятинї // „Діло“ 1889, UA-W-LVV).*

In many contexts, phrases and terms of the “noun+adjective” model that are directly borrowed from Polish are attested.

Поясненя, які я дав ***радї слїдчій****, були короткі („Діло“, 1888, UA-W-LVV). На торжество явились межи нашими* ***маршалок краєвий****, Гр. Тарновскій з членами* ***виділу краєвого****,* ***послами рускими*** *і деякими* ***послами польскими*** *(„Діло“, 1888, UA-W-LVV). Праворуч* ***престола митрополичого*** *заняли місця всі достойники* ***церкви латиньскої і вірменьскої*** *(„Діло“, 1888, UA-W-LVV). На* ***видїлї богословскім*** *єсть 11* ***професорів звичайних****; на* ***видїлї правничім****: 8* ***професорів звичайних****, 5 надзвичайних, 2 титулярні і 6* ***доцентів приватних****; на* ***видїлї философичнім****: 14* ***професорів звичайних****, 6 надзвичайних, 13* ***доцентів приватних*** *і три учителї:* ***язика Француского, англійского і стенографії*** *(Де-що з Парижа. — Львівскій университет // „Діло“ 1889, UA-W-LVV).*

The corpus also demonstrates a number of other features of the Western variant of the Ukrainian literary language up to the 1940s.

3.3. Ukrainian language of the diaspora as a modern regional variant

The linguistic phenomenon of the Western Ukrainian diaspora began to be studied after Ukraine gained independence [2, 17, etc.].

GRAC is the first and so far the only corpus of the Ukrainian language which contains texts of the Ukrainian diaspora marked by country of origin. They may be allocated to a separate subcorpus for research. The subcorpus of the diaspora in GRAC v.10 counts about 40 million tokens and contains large parts of fiction and non-fiction and a smaller share of academic literature and texts of other styles.

GRAC shows specific features in the language of the diaspora with regard to the texts of Ukraine proper, for example, the norm of declension of borrowed nouns with the final -o. In the history of the Ukrainian standard, such nouns could sometimes be declined (as in Polish) or could not be declined (as in Russian; this is the Ukrainian norm today). The following nouns belong to this class and are well represented in the corpus: *авто, бароко, бюро, вето, відео, гестапо, депо, доміно, євро, казино, какао, кіно, кредо, ласо, метро, піаніно, псевдо, радіо, рококо, соло, сопрано, танго, фортепіано, фото.*

The actual use of declined and indeclinable forms as reflected in GRAC is shown in Figure 13 and 14. They reflect the proportion of these nouns in oblique cases (e.g., *авта, авті, автом,* etc.) as opposed to all cases when they are used with final *-о* (*авто*). Cases with o include both indeclinable cases as well as regular nominative and accusative singular.

In the published texts of authors who lived permanently in Ukraine, the declension of borrowed nouns with the final *-о* was admissible until about the end of World War II. Then in the Soviet texts of 1950 – 1990 there is practically an unambiguous norm, according to which these nouns have an invariable form, just like in Russian. This is clearly reflected in the graphs.

After 1991, cases of declension of the nouns in question reappear in the texts of the corpus, but the share of this variant is considerably lower than in the 19th and first half of the 20th century. In the texts of the Ukrainian diaspora, the share of borrowings in *-о* with the endings of indirect cases in all years is larger compared to the texts written in Ukraine. However, the "declined norm" is not always consistent. See the paper [14] for more details.

This case study shows how a well known phenomenon of competing codifications can be fruitfully studied using the regional annotation of GRAC.

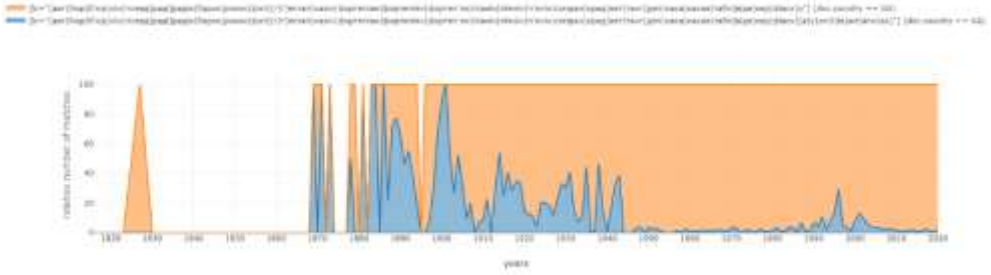

Figure 13: Borrowed nouns in *-о* with the endings of oblique cases (blue graph) and with final *-о* (orange graph) in the texts of authors who lived permanently in Ukraine.

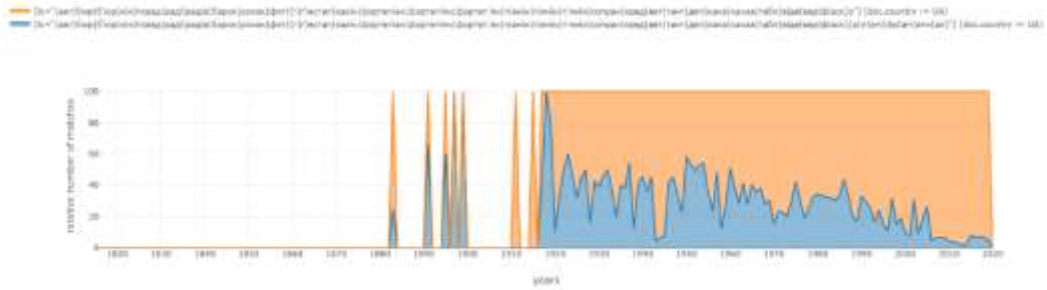

Figure 14: Borrowed nouns *-о* with the endings of indirect cases (blue graph) and with the final *-о* (orange graph) in the texts of Ukrainian authors who emigrated or resided outside Ukraine.

4. Conclusion

The regional annotation is a specific feature of GRAC. The need to add such a layer of annotation to the corpus built in the tradition of national corpora is caused by the properties and history of the Ukrainian language. The literary Ukrainian language in the 19th and early 20th centuries developed around two centers, namely Kyiv and Lviv. Later, in the 20th century, the language of the Ukrainian diaspora became yet another source for the formation of a standard norm. Nowadays, when a common literary norm has already been formed for Ukrainian as a whole, its vocabulary and grammar nevertheless remain highly varied, which is related to the history of the formation of the Ukrainian language.

GRAC thus has a layer of markup that enables comparing the language of different regions of Ukraine, as well as Ukraine and the diaspora. In the article we have shown that the corpus indeed exhibits differences in the texts of different regions, corresponding to the data of dialect maps (taking into consideration that the corpus contains written standard-oriented texts rather than dialectal) and the regional language differences known from other sources. Furthermore, we show how competing morphological norms play out differently in the diaspora and in the Ukraine, before and after 1991.

Overall, we make the case that regional variability is an important factor to be taken into account in the study of Ukrainian. GRAC as a large reference corpus of Ukrainian contains regional annotation in order to make this factor accessible and we show that this annotation can be fruitfully applied to gain insight about the current usage and norms as well as recent history of written standard Ukrainian.

References

[1] I. Matviias (Ed.), Atlas ukrainskoi movy v 3 t., Kyiv, 1988-2001. [Atlas of the Ukrainian Language in Three Volumes].

[2] B. M. Azhniuk, Movna yednist natsii: diaspora y Ukraina, Kyiv, 1999. [Language unity of the nation: diaspora and Ukraine].

[3] Ye. Baran, Leksychni hunharyzmy u tvorakh ukrainskykh pysmennykiv Zakarpattia [Lexical hungarisms in the works by the Ukrainian writers of Transcarpathia], Ukrainska mova [Ukrainian Language] 2 (2009) 56–69.

[4] N. Darchuk, Doslidnytskyi korpus ukrainskoi movy: osnovni zasady i perspektyvy [Research corpus of the Ukrainian language: basic principles and prospects], Visnyk Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka: Literaturoznavstvo, movoznavstvo, folklorystyka [Bulletin of Taras Shevchenko Kyiv National University: Studies in literature, language and folklore], 21 (2010) 45–49.

[5] O. M. Demska, Predmetna haluz zahalnomovnoho korpusu: pytannia pro surzhyk [Subject branch of the common language corpus: the question of surzhik], Naukovi Zapysky NaUKMA [Research notes of NaUKMA], 137 (2012) 17–20.

[6] O. M. Demska, Tekstovyi korpus: ideia inshoi formy, Kyiv, 2011. [Text corpus: the idea of another form].

[7] O. S. Melnychuk (Ed.), ESUM: Etymolohichnyi slovnyk ukrainskoi movy v 7 t., Kyiv, Naukova dumka, 1982-2006. [Etymological dictionary of the Ukrainian language in 7 volumes].

[8] Z. Franko, Variantnist chy terytorialna vidminnist, ukrainskoi literaturnoi movy [Variation or territorial difference of the Ukrainian literary language], Ukrainska istorychna ta dialektna leksyka [Ukrainian historical and dialectal vocabulary], 2 (1991) 169–173.

[9] M. Shvedova, R. von Waldenfels, S. Yarygin, A. Rysin, V. Starko, M. Woźniak, M. Kruk et al. GRAC: General Regionally Annotated Corpus of Ukrainian, 2017–2021, URL: http://uacorpus.org/.

[10] P. E. Hrytsenko, Nekotorye zamechanyia o dyalektnoi osnove ukraynskoho lyteraturnoho yazyka, Philologia slavica: To the 70-th anniversary of the Academician N.Y. Tolstoy (1993), 284–294. [Some remarks on the dialectal basis of the Ukrainian literary language].

[11] N. P. Darchuk, KUM: Korpus ukrainskoi movy, 2003–2021, URL: www.mova.info/corpus.aspx [Corpus of the Ukrainian language].

[12] I. H. Matviias, Vzaiemodiia skhidnoukrainskoho y zakhidnoukrainskoho variantiv literaturnoi movy v ustalenni norm u haluzi syntaksysu [Interaction of East Ukrainian and West Ukrainian

versions of literary language in establishing norms in the field of syntax], Movoznavstvo [Linguistics], 1 (2013) 3–8.
[13] I. Matviias, Varianty ukrainskoi literaturnoi movy v kintsi XVIII i v XIX stolitti [Variants of the Ukrainian literary language in the late 18-th and 19-th centuries], Kultura slova [Culture of the Word] 48-49 (1996) 11–28.
[14] M. O. Shvedova, Hramatychne osvoiennia zapozychenykh imennykiv iz kintsevym -o v ukrainskii movi: korpusne doslidzhennia [Grammatical mastering of borrowed nouns with the final -o in the Ukrainian language: corpus research], Ukrainska mova [Ukrainian language], 2 (2020) 13–30.
[15] M. Shvedova, 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, Vol. I: Main Conference. Lviv, Ukraine, April 23-24, 2020, pp. 489–506.
[16] M. Shymkova, Reprezentatyvnost korpusa kak lynhvystycheskaia problema [Corpus representativeness as a Linguistic Problem], Tr. Mezhdunar. konf. MegaLing'2005 [Proceedings of the International Conference MegaLing'2005], Saint Petersburg, Osypov, 2005, pp. 130–139.
[17] O. O. Taranenko, Mova ukrainskoi zakhidnoi diaspory i suchasna movna sytuatsiia v Ukraini (na zahalnoslov'ianskomu tli) [The language of the Ukrainian Western Diaspora and the current language situation in Ukraine (against the Slavic background)], Movoznavstvo [Linguistics], 2-3 (2013) 63-99.
[18] P. Auer, Dialect vs. standard: a typology of scenarios in Europe, In: B. Kortmann, J. var der Auwera (Eds.), The Languages and Linguistics of Europe: A Comprehensive Guide, De Gruyter Mouton, Germany, 2011, pp. 485–500.
[19] D. Biber, Representativeness in corpus design, Literary and linguistic computing, 8(4), (1993) 243–257.
[20] J. Chromý, Korpus a reprezentativnost, Naše řeč, ročník, 97 (2014) 185–193. URL: http://nase-rec.ujc.cas.cz/archiv.php?art=8337.
[21] J. Sinclair, Reference Corpora, EAGLES Preliminary recomendations on Corpus Typology, 1996, URL: http://www.ilc.cnr.it/EAGLES96/corpustyp/node18.html [Accessed on 10.04.2021].