Corpus-driven Approach to Ukrainian E-anecdotes Study

Alla Luchyk[a], Oksana Taran[b], Oleksandra Palchevska[c], Natalia Sharmanova[d], Ganna Demydenko[d]

[a] *National University of "Kyiv-Mohyla academy", Skovoroda Street 2, Kyiv, 04655, Ukraine*
[b] *Lviv National Polytechnik University, Bandera Street 12, Lviv, 79013, Ukraine*
[c] *Lviv State University of Life Safety, Kleparivska Street 35, Lviv, 79007, Ukraine*
[d] *Kryvyi Rih State Pedagogical University, Gagarin Avenue 54, Kryvyi Rih, 50086, Ukraine*

Abstract

The Ukrainian e-anecdotes corpus was created in order to describe structural and linguistic features of Ukrainian e-anecdotes. It contains 500 anecdotes, its volume is 18,582 tokens. Corpus-driven approach with Sketch Engine corpus management allowed to describe some linguistic and quantitative characteristics of Ukrainian e-anecdotes, to identify rare/unusual words in the corpus of e-anecdotes and interpret them, to analyze the keyword collocations and to determine linguistic and genre features of Ukrainian anecdotes, as well as to identify the ethnonymic collocations as the ethnic stereotypes markers.

Keywords [1]

Corpus, the Ukrainian language e-anecdote, frequency, score, Skecth Engine

1. Introduction

Despite the fact that since the appearance of the term “anecdote” in 550 in Byzantium (the historians P. Caesarea’s book “Secret History”), many works have been written, the anecdotes genre and linguistic features study in different languages continues. Ukrainian language studies are not rich in research: Ukrainian anecdotes are considered from the linguistic, folklore, literary studies, ethnopsychology point of view [1, 2, 3, 4, 5, 6], the Internet anecdote (e-anecdote) as a genre becomes more actual during the last decade [7, 8, 9, 10, 11], however, the studies of the kind on the Ukrainian language material do not exist.

We have analyzed all the open corpora developed for the Ukrainian language (section 2.1), but none of them contains the anecdotes’ subcorpus, and accordingly, there are no anecdotes’ corpus-based studies, what emphasizes the urgency of this work.

The purpose of our article is to create a corpus of e-anecdotes and to show the corpus capabilities in the study as well as the structural and linguistic features of Ukrainian e-anecdotes through their quantitative parameters prism. This corpus-driven approach is used in the study, which is aimed to interpret corpus data as a whole [12].

First we will present the existing resources and NLP tools developed for the Ukrainian language or those that can be used (section 2) for this purpose, and then introduce the Ukrainian e-anecdotes corpus (section 3) as well its exploitation for the Ukrainian e-anecdotes linguistic and quantitative characteristics descrition (section 4).

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine

EMAIL: allal@meta.ua (A. Luchyk); oksana.s.taran@lpnu.ua (O. Taran); palch56@ukr.net (O. Palchevska); nmsharm@gmail.com (N. Sharmanova); demidenko_anna@ukr.net (G. Demydenko);
ORCID: 0000-0002-4218-355X (A. Luchyk); 0000-0001-6312-9563 (O. Taran); 0000-0002-2090-4870 (O. Palchevska); 0000-0003-4820-3619 (N. Sharmanova); 0000-0003-1494-2536 (G. Demydenko)

2. Existing resources and tools

First we offer a short review of the open Ukrainian corpora and some NLP tools.

2.1. Ukrainian corpora

Ukrainian web corpus of the University of Leipzig (2019, 1 billion tokens) [13]. In the search results there are the examples, compatibility, three-dimensional connectivity graph. However, the search is possible only by the word form.

Ukrainian language corpus of the "Chtivo" library (600 million words, 6.6 GB of Ukrainian texts from the electronic "Chtivo" library) [14] – these are automatically recognized fiction, scientific, journalistic texts, that do not have the errors correction and corpus annotation.

"Ukrainian laboratory" [15] consists of: *web body of Internet texts* with syntactic markup of almost 3 billion tokens*, "Golden corpus"* – it is a manually annotated corpus of the various styles texts with removed homonymy (140 thousand tokens); *parallel corpora of fiction* (6 million tokens).

Corpora of Ukrainian texts (authors: Dmytro Chaplynsky and Vsevolod Dyomkin) [16] contain *The corpus of NER-annotations* (has 229 texts from the Ukrainian Brownian corpus of 217,381 tokens from 6,751 annotated named entities), *Corpus UberText* that is a corpus of Ukrainian periodicals (over 6 GB) where Andriy Rysin and BRUK initiative nlp-uk package was used for texts tokenization and lemmatization. *Corpus of laws and legal acts* – corpus (over 9 GB) with tokenization and lemmatization as well as the Word Embedding models.

Corpus of the Ukrainian language on the linguistic portal Mova.info (2003–2020) [17] developed in the computer linguistics laboratory of the Institute of Philology of Taras Shevchenko National University of Kyiv under the prof. N. Darchuk direction. The corpus contains the following subcorpora: legislative texts, scientific texts, poetic language, journalism, folklore texts, fiction. It has great functionality.

General Regionally Annotated Corpus of Ukrainian (GRAC) [18] – corpus of texts annotated by regions, containing more than 650 million tokens (v. 10). Authers: M. Shvedova, R. Von Waldenfels, S. Yarygin, A. Rysin, V. Starko, M. Wozniak, M. Kruk and others. The corpus covers the period from 1816 to 2020 and contains more than 900,000 texts by about 26,000 authors. Texts have the following meta-marking: by styles, themes, genres, original language, dating, spelling, information about the author of the text, information about the media, by regions, by the text source. The humor is presented among the genres (search by tag: <doc genre="HUM"/>), but there no anecdotes among the texts. The corpus has a linguistic annotation: morphological (works on the morphological analysis system basis developed by the r2u group specialists Andriy Rysin, Vasyl Starko and others), grammatical homonymy (removal of homonymy is done manually, but inconsistently), semantic for 3000 most frequent words [19, 20].

Grammatical Error Correction and Fluency Corpus for the Ukrainian Language [21] – this is a new and ambitious project from Grammarly. Now UA-GEC containes 328 779 tokens.

The Ukrainian Brown Corpus [22] is under development (authors: Vasyl Starko, Andriy Rysin and others). Three more corpora, that are under development, can be mentioned: *The Medical Corpus in Ukrainian (*UKRMED*)* [23], *The multilingual aligned corpus with Ukrainian as the target language* [24] and *Ukrainian corpus WikiWars-UA* [25]. The last one is annotated with temporal information and also proposes a comparison with the English version of annotations.

2.2. NLP tools

To perform the objectives of this study, we are going to analyze the free lexicons, PoS-taggers and parsing tools developed for the Ukrainian language or another Slavic inflectional language, useful for our work.

MULTEXT-East free lexicons 4.0 is a morphosyntactic lexicons resource which also contains Ukrainian lexicon (300292 entries) [26], which can be used in AntConc for Keyword List.

Mystem+ is an open source of automatic morphological analysis for Russian such as: *TreeTagger*, *Mystem*, TnT, *Hunpos and* MarMoT. There are also exist the comparative study of automatic morphological analysis for Russian [27]. Asiryan A. K. described a comparison of morphological analyzers *TreeTagger*, *MyStem*, *TnT*, *pymorphy2* and *FreeLing* for three Russian corpora [28].

FreeLing is an open source language analysis tool suite providing language analysis functionalities (morphological analysis, named entity detection, PoS-tagging, parsing, Word Sense Disambiguation, Semantic Role Labelling, etc.) for a variety of languages (English, Spanish, Portuguese, Italian, French, German, Russian, Catalan, Galician, Croatian, Slovene , among others) [29]. Ukrainian is not on the list, however FreeLing version 4.2 copes well with parsing as for the Russian language texts.

The greatest interest for the anecdotes' dialogic speech analysis is *Dialogue Annotation and Research Tool (DART)* Version 3.0, author: Martin Weisser (May 2019). It is a research environment designed to allow you to annotate and analyse single or multiple dialogues in batch mode, with the ultimate aim to identify speech acts automatically, and thereby create pragmatically annotated corpora. The version 3 identifies 162 speech acts automatically, and also has a number of additional functions and improvements to both the interface and the output options from within the individual analysis modules, as well as a completely new pattern counting facility [30]. The DART Taxonomy version 3.0 is designed for English and consist of 2 parts: speech-act label and function.

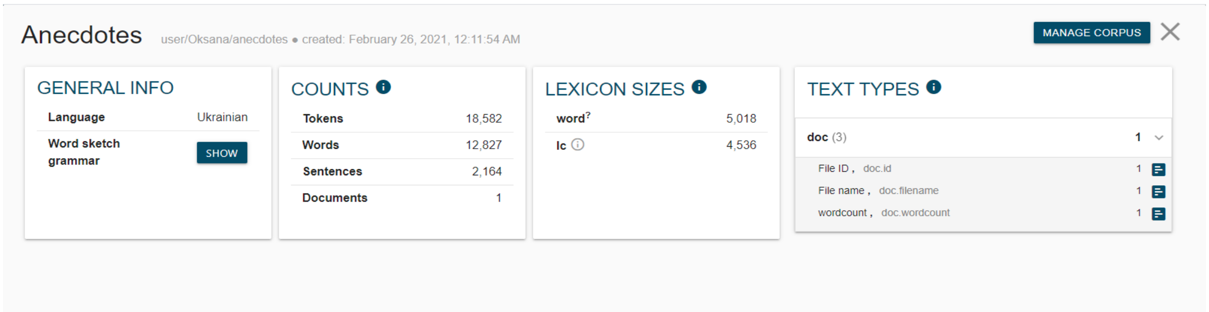
Project to generate POS tag dictionary for Ukrainian language. For all files in data/dict the project generates all possible word forms with POS tags by using affix rules from files in data/affix [31].

AntConc [32] is a freeware corpus analysis toolkit for concordancing and text analysis. It is good for Ukrainian but we had to append the Ukrainian alphabet to Token Definition Settings.

In reality, there exist only few resources for studying the Ukrainian language. The most suitable for our tasks is *Sketch Engine* [33]. It is a commercial corpus management and corpus query software with user friendly interface. It has no PoS-tagging for Ukrainian but can perform parsing.

3. Ukrainian e-anecdotes corpus

Collection of Ukrainian e-anecdotes is composed of 500 texts by random sampling from various Ukrainian-language sites. All texts were normalized (for example, the different codes use in a set of words is detected *тама≠мама, hiяк≠ніяк, hа≠на (mother*, *any*, *take)*, typing errors (*розмляють* instead of *розмовляють/talk*). Further corpus exploitation is done with the Sketch Engine tools. Our corpus statistics is given on the Figure 1.



Anecdotes user/Oksana/anecdotes • created: February 26, 2021, 12:11:54 AM

GENERAL INFO

Language	Ukrainian
Word sketch grammar	SHOW

COUNTS

Tokens	18,582
Words	12,827
Sentences	2,164
Documents	1

LEXICON SIZES

word?	5,018
lc	4,536

TEXT TYPES

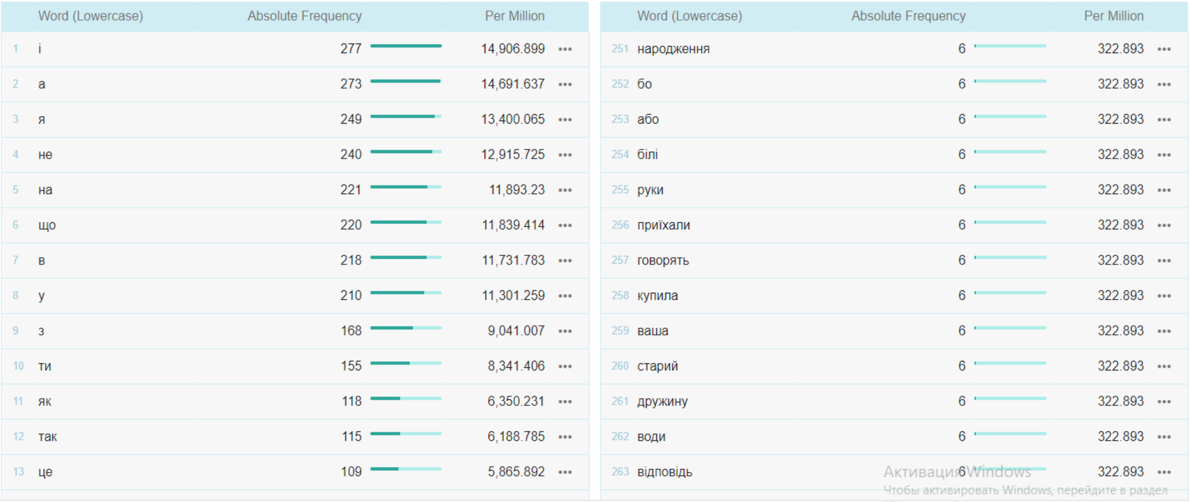
doc (3)	1
File ID , doc.id	1
File name , doc.filename	1
wordcount , doc.wordcount	1

Figure 1: Corpus statistics

4. Exploitation of corpus and results of the study

4.1. Some linguistic and quantitative characteristics of the e-anecdotes corpus

The WordList option allows you to find the most frequent tokens in anecdotes (and further lemmatization will allow you to identify lemmas). According to A. Shmelev, an anecdote is a formal dialogue, because at the beginning the sender asks a question, but without waiting for the answer from the addressee, he answers it himself [34]. The frequency list (Figure 2) shows the high frequency of personal pronouns different forms (*я, ти, мене, він, ви, мені, його, тебе/ I, you, me, he, you, me him, you*). This is in line with the dialogue format.



	Word (Lowercase)	Absolute Frequency	Per Million
1	і	277	14,906.899
2	а	273	14,691.637
3	я	249	13,400.065
4	не	240	12,915.725
5	на	221	11,893.23
6	що	220	11,839.414
7	в	218	11,731.783
8	у	210	11,301.259
9	з	168	9,041.007
10	ти	155	8,341.406
11	як	118	6,350.231
12	так	115	6,188.785
13	це	109	5,865.892

	Word (Lowercase)	Absolute Frequency	Per Million
251	народження	6	322.893
252	бо	6	322.893
253	або	6	322.893
254	білі	6	322.893
255	руки	6	322.893
256	приїхали	6	322.893
257	говорять	6	322.893
258	купила	6	322.893
259	ваша	6	322.893
260	старий	6	322.893
261	дружину	6	322.893
262	води	6	322.893
263	відповідь	6	322.893

Figure 2: WordList of e-anecdotes corpus

Parsing allows to determine the sentences quantitative parameters by intonation (Table 1).

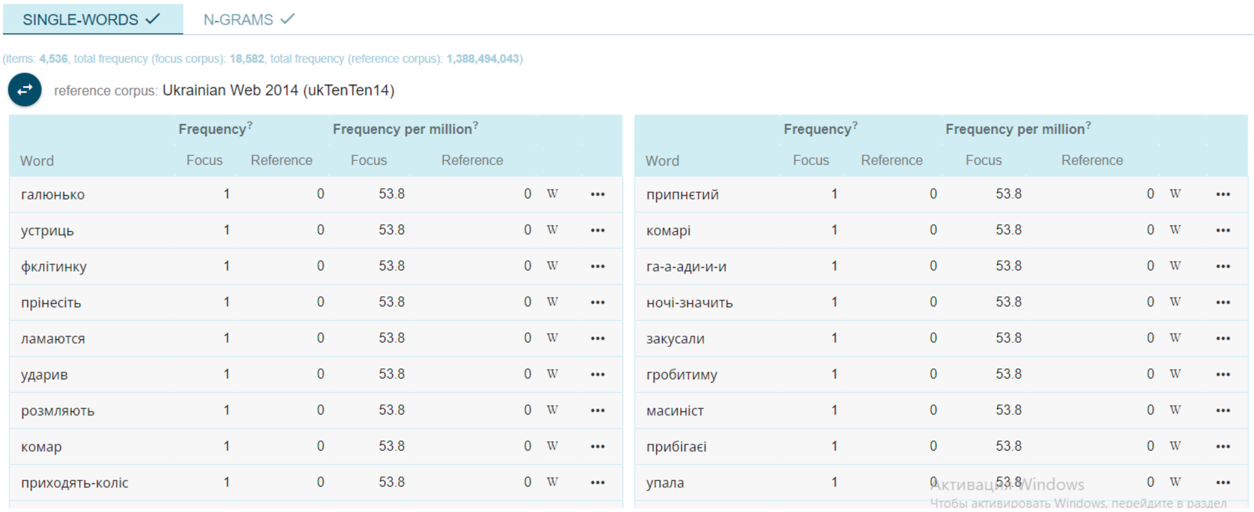
Table 1
Sentences quantitative parameters according to intonation design

Types of sentences	Declarative	Interrogative	Exclamatory	Interrogative and exclamatory	Ellipticali	Pseudo-sentences
Absolute frequency	1007	560	427	24	138	8
Relative frequency, %	46,5	25,9	19,7	1,1	6,4	0,4

Pseudo-sentences consist only of punctuation marks (an ellipsis or several exclamation marks) and do not contain verbal components. As you can see, half of the sentences in the corpus are intonationally and emotionally colored.

4.2. Determination of rare or unusual words in the e-anecdotes corpus

Sketch Engine comprises only one Ukrainian corpus. It is Ukrainian Web 2014 (ukTenTen14) amounting to 1,388,494,043 tokens. So, it can reflect the general language trends. It was chosen as the reference corpus. The anecdotes corpus is a focus corpus. If a word with frequency 1 in focus corpus has frequency 0 in the reference corpus, it is considered as rare or unusual word (Figure 3). In order to perform this task advanced settings were used.



SINGLE-WORDS ✓ N-GRAMS ✓

(items: **4,536**, total frequency (focus corpus): **18,582**, total frequency (reference corpus): **1,388,494,043**)

reference corpus: Ukrainian Web 2014 (ukTenTen14)

	Frequency?		Frequency per million?			
Word	Focus	Reference	Focus	Reference		
галюнько	1	0	53.8	0	W	•••
устриць	1	0	53.8	0	W	•••
фклітинку	1	0	53.8	0	W	•••
прінесіть	1	0	53.8	0	W	•••
ламаются	1	0	53.8	0	W	•••
ударив	1	0	53.8	0	W	•••
розмляють	1	0	53.8	0	W	•••
комар	1	0	53.8	0	W	•••
приходять-коліс	1	0	53.8	0	W	•••

	Frequency?		Frequency per million?			
Word	Focus	Reference	Focus	Reference		
припнєтий	1	0	53.8	0	W	•••
комарі	1	0	53.8	0	W	•••
га-а-ади-и-и	1	0	53.8	0	W	•••
ночі-значить	1	0	53.8	0	W	•••
закусали	1	0	53.8	0	W	•••
гробитиму	1	0	53.8	0	W	•••
масиніст	1	0	53.8	0	W	•••
прибігаєі	1	0	53.8	0	W	•••
упала	1	0	53.8	0	W	•••

Figure 3: Rare or unusual words in the corpus

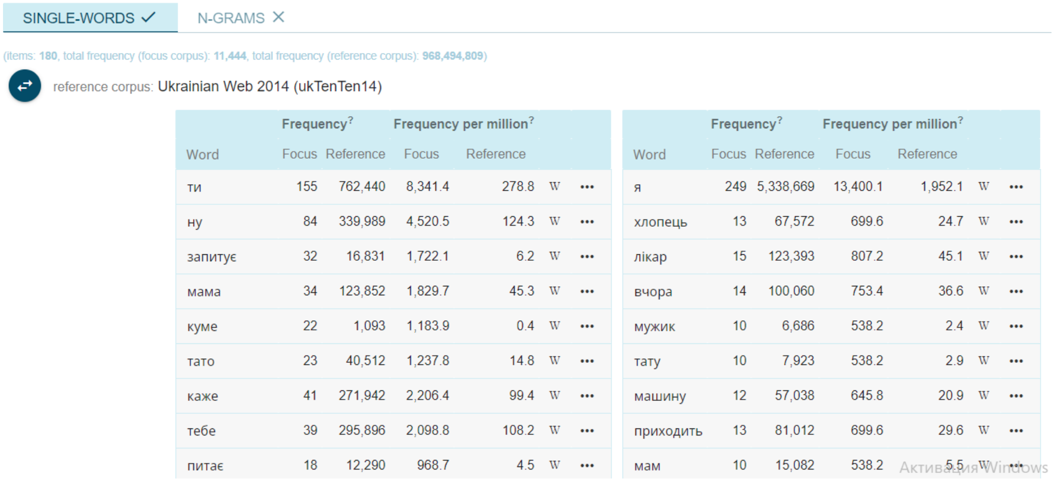
As a result, 997 such tokens (5%) were detected. It:

- surzhykisms (*валідольчік, балуєшся, лавірував/validol, roistering, showed dexterity*)
- dialectisms (*Йсусу, вдівся/ Jesus, dressed*)
- slang vocabulary (*крякнути/to crack*)
- graphically unstable neologisms (*смска, смс-ку/sms*)
- the accusative case forms (*Романівно, доцю, куме, Русланчику/Romanivno, daughter, crony, Ruslanchiku*)
- Ukrainian graphic forms of Russian words (*московскоє/moskow*)
- specific exclamations (*уiiii, шшшшш/ yiiii,shshshshsh*)
- the reproduction of emphasis (тааат, га-а-ади-и-и, мааам, о-о-ооо, уф-ф-ф, таак/dadyyy, ga-a-adi-i-i, maaam, o-o-ooo, uf-f-f, yeees)
- the reproduction of the character's speech in a state of altered consciousness (for example: *А коли ти зрозумів, що вона п'яна? — Коли прийшла СМС "**Повзони** мені, а то я не можу найти свій **мефелоон**" / And when did you realize that she was drunk? — When the SMS came "**Coull** me, and for I can not find my **tefeloon**")*
- the reproduction of the character's distorted articulation (*фклітинку/flaid* – children's speech; *мяфко, фамо, прифкакало/meat, itself, jumped* – consonant changes because of a stuffed mouth, *дддома, сість, масиніст/at home, six, train draiver* – due to physical condition)
- the reproduction of Ukrainian words pronunciation by a Russian-speaking character (*хлопци, останівка/ guys, stop*)
- barbarisms (*lakalut, dont*)
- words used in a certain grammatical form only one time (*комарів, устриць/mosquitoes, oysters*).

Thus, the anecdote as a conversational genre reflects the full oral non-formal speech linguistic means range.

4.3. KeyWords in the corpus

With the same tool but on the different criteria words which are very frequent in the focus corpus were determined, what allows to compare with the frequency in the referent corpus (Figure 4). In such case it is relevant to compare the frequency per million.



(items: 180, total frequency (focus corpus): 11,444, total frequency (reference corpus): 968,494,809)

reference corpus: Ukrainian Web 2014 (ukTenTen14)

	Frequency?		Frequency per million?			
Word	Focus	Reference	Focus	Reference		
ти	155	762,440	8,341.4	278.8	W	•••
ну	84	339,989	4,520.5	124.3	W	•••
запитує	32	16,831	1,722.1	6.2	W	•••
мама	34	123,852	1,829.7	45.3	W	•••
куме	22	1,093	1,183.9	0.4	W	•••
тато	23	40,512	1,237.8	14.8	W	•••
каже	41	271,942	2,206.4	99.4	W	•••
тебе	39	295,896	2,098.8	108.2	W	•••
питає	18	12,290	968.7	4.5	W	•••

	Frequency?		Frequency per million?			
Word	Focus	Reference	Focus	Reference		
я	249	5,338,669	13,400.1	1,952.1	W	•••
хлопець	13	67,572	699.6	24.7	W	•••
лікар	15	123,393	807.2	45.1	W	•••
вчора	14	100,060	753.4	36.6	W	•••
мужик	10	6,686	538.2	2.4	W	•••
тату	10	7,923	538.2	2.9	W	•••
машину	12	57,038	645.8	20.9	W	•••
приходить	13	81,012	699.6	29.6	W	•••
мам	10	15,082	538.2	5.5	W	•••

Figure 4: Common frequent words

Comparison of the sorting by frequency results in the focus corpus and separately in the reference corpus revealed that in the reference corpus the first 18 positions are occupied by prepositions and conjunctions, and in the focus corpus by prepositions, conjunctions, personal pronouns and exclamations. Thus, the anecdote as a conversational genre is characterized by a high frequency of expressive means and personal pronouns, the use of which creates the narrator-eyewitness effect for the events depicted in the anecdote.

In the obtained result, all nouns (N) and verbs (V) with a frequency of 10 were marked with the help of MS Excel tools, semantically close (*запитує=питає/asks)* were lemmatized and combined. Table 2 represents the most frequent nouns and verbs list.

Table 2

Absolute frequency of nouns and verbs lemmas

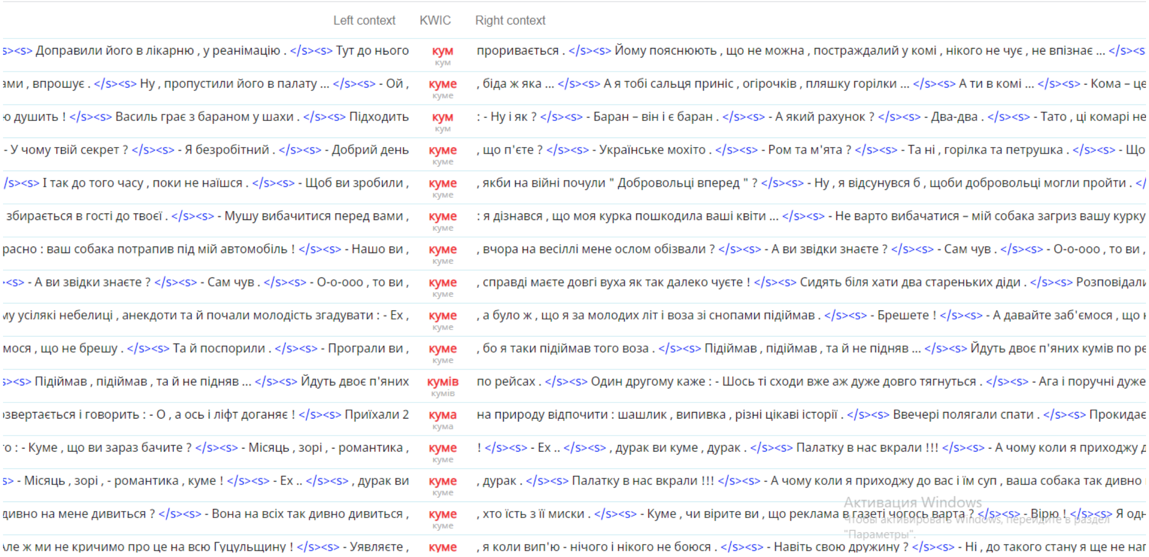
Part of speech	Lemma	Absolute frequency	Part of speech	Lemma	Absolute frequency
N	"мама"	55	V	"запитувати"	50
N	"чоловік"	49	V	"говорити"	30
N	"тато"	33	V	"бачити"	17
N	"хлопець"	32	V	"приходити"	13
N	"дружина"	24	V	"кричати"	12
N	"кум"	22	V	"працювати"	12
N	"жінка"	21	V	"розмовляти"	12
N	"день"	16	V	"знати"	11
N	"лікар"	15	V	"сидіти"	11
N	"син"	14	V	"вирішити"	10
N	"блондинка"	13	V	"хотіти"	10
N	"доктор"	12			
N	"машина"	12			
N	"рік"	12			

N	"ведмідь"	11
N	"дівчина"	11
N	"очі"	11
N	"студент"	11
N	"дитина"	10
N	"кіт"	10
N	"мужик"	10
N	"професор"	10
N	"українець"	10
N	"час"	10

The anecdote is of an exclusively anthropocentric direction. Ukrainian anecdotes are thematically diverse: anecdotes about the family, friends and godparents, neighbors, school and university, professions, about Vovochka (they became especially relevant due to an allusion to the leader of the Russian Federation), personalized animals, gender roles, different ethnic groups. The most frequent keywords indicate the anecdotes about family and gender predominance in the corpus.

4.4. Concordans and collocations

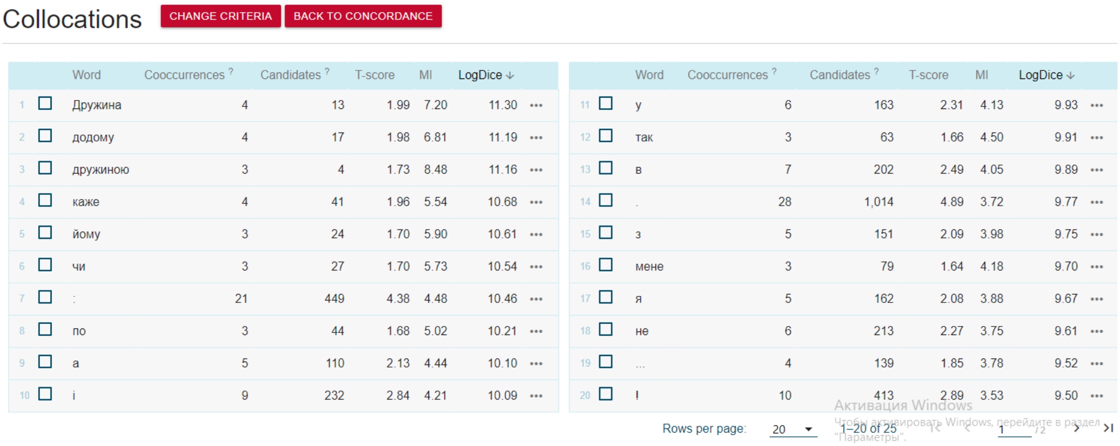
The use of concordance for the most frequency keywords made it possible to identify some linguistic features. For example, the predominance of the accusative case for the token *кум/crony* (Figure 5). Accordingly, it appears in the speech of the characters.



Left context	KWIC	Right context
s><s> Доправили його в лікарню , у реанімацію . </s><s> Тут до нього	кум кум	проривається . </s><s> Йому пояснюють , що не можна , постраждалий у комі , нікого не чує , не впізнає ... </s><s
ами , впрошує . </s><s> Ну , пропустили його в палату ... </s><s> - Ой ,	куме куме	, біда ж яка ... </s><s> А я тобі сальця приніс , огірочків , пляшку горілки ... </s><s> А ти в комі ... </s><s> - Кома – це
ю душить ! </s><s> Василь грає з бараном у шахи . </s><s> Підходить	кум кум	: - Ну і як ? </s><s> - Баран – він і є баран . </s><s> - А який рахунок ? </s><s> - Два-два . </s><s> - Тато , ці комарі не
- У чому твій секрет ? </s><s> - Я безробітний . </s><s> - Добрий день	куме куме	, що п'єте ? </s><s> - Українське мохіто . </s><s> - Ром та м'ята ? </s><s> - Та ні , горілка та петрушка . </s><s> - Що
/s><s> І так до того часу , поки не наїшся . </s><s> - Щоб ви зробили ,	куме куме	, якби на війні почули " Добровольці вперед " ? </s><s> - Ну , я відсунувся б , щоби добровольці могли пройти . </
збирається в гості до твоєї . </s><s> - Мушу вибачитися перед вами ,	куме куме	: я дізнався , що моя курка пошкодила ваші квіти ... </s><s> - Не варто вибачатися – мій собака загриз вашу курку
расно : ваш собака потрапив під мій автомобіль ! </s><s> - Нашо ви ,	куме куме	, вчора на весіллі мене ослом обізвали ? </s><s> - А ви звідки знаєте ? </s><s> - Сам чув . </s><s> - О-о-ооо , то ви ,
<s> - А ви звідки знаєте ? </s><s> - Сам чув . </s><s> - О-о-ооо , то ви ,	куме куме	, справді маєте довгі вуха як так далеко чуєте ! </s><s> Сидять біля хати два старенько діди . </s><s> Розповідали
му усілякі небелиці , анекдоти та й почали молодість згадувати : - Ех ,	куме куме	, а було ж , що я за молодих літ і воза зі снопами підіймав . </s><s> - Брешете ! </s><s> - А давайте заб'ємося , що і
мося , що не брешу . </s><s> Та й поспорили . </s><s> - Програли ви ,	куме куме	, бо я таки підіймав того воза . </s><s> Підіймав , підіймав , та й не підняв ... </s><s> Йдуть двоє п'яних кумів по ре
s><s> Підіймав , підіймав , та й не підняв ... </s><s> Йдуть двоє п'яних	кумів кумів	по рейсах . </s><s> Один другому каже : - Шось ті сходи вже аж дуже довго тягнуться . </s><s> - Ага і поручні дуже
звертається і говорить : - О , а ось і ліфт доганяє ! </s><s> Приїхали 2	кума кума	на природу відпочити : шашлик , випивка , різні цікаві історії . </s><s> Ввечері полягали спати . </s><s> Прокидає
о : - Куме , що ви зараз бачите ? </s><s> - Місяць , зорі , - романтика ,	куме куме	! </s><s> - Ех .. </s><s> , дурак ви куме , дурак . </s><s> Палатку в нас вкрали !!! </s><s> - А чому коли я приходжу д
s> - Місяць , зорі , - романтика , куме ! </s><s> - Ех .. </s><s> , дурак ви	куме куме	, дурак . </s><s> Палатку в нас вкрали !!! </s><s> - А чому коли я приходжу до вас і їм суп , ваша собака так дивно
дивно на мене дивиться ? </s><s> - Вона на всіх так дивно дивиться ,	куме куме	, хто їсть з її миски . </s><s> - Куме , чи вірите ви , що реклама в газеті чогось варта ? </s><s> - Вірю ! </s><s> Я одн
але ж ми не кричимо про це на всю Гуцульщину ! </s><s> - Уявляєте ,	куме	, я коли вип'ю - нічого і нікого не боюся . </s><s> - Навіть свою дружину ? </s><s> - Ні , до такого стану я ще не нап

Figure 5: Concordance for keyword *кум/crony*

For each concordance we verified collocations with such criteria as: range from –7 to 7, custom range, minimum frequency in corpus is 1, minimum frequency in a given range is 3. It allows to reveal paired characters: *дружина* — *чоловік, мама* — *тато/wife – husband, mom – dad*, etc., since these tokens have the highest LogDice as a statistic score for identifying collocations. Figure 6 shows the collocations search results for the lexeme *чоловік/husband.*



Collocations

		Word	Cooccurrences ?	Candidates ?	T-score	MI	LogDice ↓	
1	□	Дружина	4	13	1.99	7.20	11.30	•••
2	□	додому	4	17	1.98	6.81	11.19	•••
3	□	дружиною	3	4	1.73	8.48	11.16	•••
4	□	каже	4	41	1.96	5.54	10.68	•••
5	□	йому	3	24	1.70	5.90	10.61	•••
6	□	чи	3	27	1.70	5.73	10.54	•••
7	□	:	21	449	4.38	4.48	10.46	•••
8	□	по	3	44	1.68	5.02	10.21	•••
9	□	а	5	110	2.13	4.44	10.10	•••
10	□	і	9	232	2.84	4.21	10.09	•••
11	□	у	6	163	2.31	4.13	9.93	•••
12	□	так	3	63	1.66	4.50	9.91	•••
13	□	в	7	202	2.49	4.05	9.89	•••
14	□	.	28	1,014	4.89	3.72	9.77	•••
15	□	з	5	151	2.09	3.98	9.75	•••
16	□	мене	3	79	1.64	4.18	9.70	•••
17	□	я	5	162	2.08	3.88	9.67	•••
18	□	не	6	213	2.27	3.75	9.61	•••
19	□	...	4	139	1.85	3.78	9.52	•••
20	□	!	10	413	2.89	3.53	9.50	•••

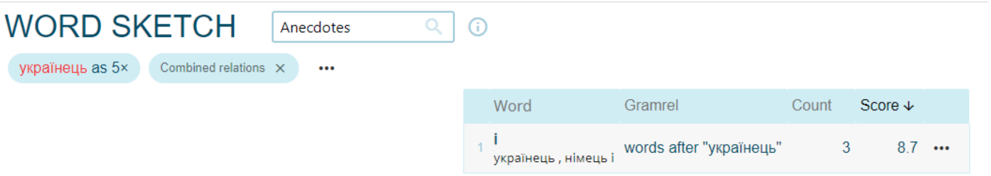
Figure 6: Collocations with the word *чоловік/husband* in the corpus of e-anecdotes

4.5. Collocation with ethnonyms as ethnic stereotypes markers

Stereotypes are the different assumptions and standardized norms put by the society on people belonging to certain groups, based on their gender, race, ethnicity or age etc. These opinions or assumptions do not correspond to the truth as they discriminate the individuals without any regard for their feelings or emotions [35]. There are exist stereotypes of different nature: sexual, gender, racist, ethnic, group.

V. Samokhina distinguishes between two types of ethnic jokes: that of the first type describe one ethnic group, of the second – several. The latter often depict other ethnic groups on the negative side [36]. In addition, the factor whether a person characterizes himself/herself (autostereotype) or another ethnic group or ethnic groups (heterostereotype) should be considered.

The ethnonym *українець/Ukrainian* was found among the frequency KeyWords. This indicates the predominance in the corpus of anecdotes with the autostereotype, however, perhaps in contrast to other ethnic groups. In order to find out with which ethnocharacters Ukrainians most often correlate in anecdotes, we have used the Typically Score (Figure 7). It shows how strong the collocation is. Obtained results for minimum score 5 showed only one case:

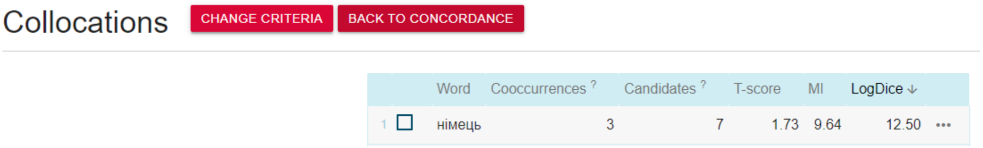


WORD SKETCH

	Word	Gramrel	Count	Score ↓	
1	і українець , німець і	words after "українець"	3	8.7	•••

Figure 7: Typically Score for word *українець/Ukrainian*

Figure 8 shows the association ratio (MI), the frequency of the keyword and the collocation in the text co-occurrence (T-score) in 2-Grams. For calculating of these scores formulas see: [37].



Collocations

		Word	Cooccurrences ?	Candidates ?	T-score	MI	LogDice ↓	
1	□	німець	3	7	1.73	9.64	12.50	•••

Figure 8: Collocations with the ethnonym *українець/Ukrainian* in the corpus of e-anecdotes

Simple search in Concordance allows to find in anecdotes the following ethno-characters, which are compared with Ukrainians: *грузин, китаєць, японець, француз/Georgian, Chinese, Japanese, French.* Anecdotes about one ethnic group present the following ethnic characters: *американець, єврей, москаль/American, Jew, Muscovite.*

5. Conclusions and future work

Created Ukrainian e-anecdotes corpus contains 500 anecdotes, body volume 18 582 tokens, parsing is done by sentences. Corpus-driven approach with Sketch Engine tools allowed to identify the following linguistic and genre characteristics of the anecdote within the corpus:

- various forms personal pronouns' high frequency, due to the anecdotes' dialogic form
- in terms of intonation, about 50% of sentences are intonationally and emotionally colored
- anecdote as a conversational genre reflects the full range of oral non-formal speech linguistic means
- KeyWords in the corpus testify the anthropocentric direction of anecdotes and the most frequent keywords indicate the family and gender anecdotes predominance in the corpus
- analysis of collocations by LogDice of keywords revealed some features: the predominance of the accusative case for the lexeme *кум/crony*, paired characters: *wife – husband, mother – father*, collocations of the ethnonym *Ukrainian* in the e-anecdotes corpus correlates with the word *German*.

So as GRAC is the largest in volume and functionality corpus and gives the possibility to create your own subcorpus of selected texts, it was chosen to download the created Ukrainian e-anecdotes corpus for future research (see GRAC-11). This is the practical significance of this work on creating the e-anecdotes corpus.

As the GRAC corpus has a semantic meta-markup that gives the possibility for proper names search, the onym space of Ukrainian anecdotes can be explored in the future.

6. References

[1] O. O. Stokolos-Voronchuk, Etnotypy ta yikh khudozhni stereotypy v ukrayinsʹkomu anekdoti (na materiali vydanʹ XIX st.) [Ethnotypes and their fiction stereotypes in the Ukrainian anecdote (on the material of publications of the XIX century)], Lviv, 2005. (in Ukrainian)

[2] I. I. Kimakovych, Folʹklornyy anekdot yak zhanr: monohrafiya [Folklore anecdote as a genre: a monograph], Kiyiv, 2006. (in Ukrainian)

[3] R. Kyrchiv, Etyudy do studiy nad ukrayinsʹkym narodnym anekdotom [Etudes for studies on the Ukrainian folk anecdote], Lviv, 2008. (in Ukrainian)

[4] L. Kornyeva, Pro deyaki typolohichni rysy ukrayinsʹkoho anekdotu [On some typological features of the Ukrainian anecdote], Philologichni nauki. 1 (2009) 55–61. URL: http://nbuv.gov.ua/UJRN/Fil_Nauk_2009_1_11

[5] N. Holovetsʹka, Motyvy ta poetyka ukrayinsʹkoho politychnoho anekdotu 90-kh rokiv XX stolittya [Motives and poetics of the Ukrainian political anecdote of the 90s of the XX century], Literatura. Folklor. Problemy poetyky. 37.1 (2012) 64 –72. (in Ukrainian)

[6] O. Kalyta, Typy komichnykh tekstiv ta osoblyvosti yikh stylistychnoho analizu [Types of comic texts and features of their stylistic analysis], Linhvostatystychni studiyi. 2 (2015) 54–61. (in Ukrainian)

[7] T. J. Blank, Toward a Conceptual Framework for the Study of Folklore and the Internet, in: Trevor J. Blank (Ed.) Folklore and the Internet: Vernacular Expression in a Digital World, Utah State University Press, Logan, 2009, p. 9.

[8] M. D. Alekseyevskiy, Internet v fol'klore ili fol'klor v Internete? (Sovremennaya fol'kloristika i virtual'naya real'nost') [Internet in folklore or folklore on the Internet? (Contemporary folklore and virtual reality)], in: Ot kongressa k kongressu: Navstrechu Vtoromu Vserossiyskomu kongressu fol'kloristov, Moscow, 2010, pp. 151–166. (in Russian)

[9] L. Shifman, H. Levy, M. Thelwall, Internet Jokes: The Secret Agents of Globalization?, Journal of Computer-Mediated Communication, vol. 19, issue 4, 1 July 2014, pp. 727–743. URL: https://doi.org/10.1111/jcc4.12082

[10] L. Boxman-Shabtai, L. Shifman, Digital humor and the articulation of locality in an age of global flows, HUMOR, vol. 29, no. 1, 2016, pp. 1–24. URL: https://doi.org/10.1515/humor-2015-0127

[11] V. V. Dement'yev, Internet-anekdoty: nekotoryye strukturnyye tipy [Internet jokes: some structural types], Zhanry rechi, 1 (15), 2017, pp. 118–135. doi: 10.18500/2311-0740-2017-1-15-118-135. (in Russian)

[12] E. Tognini-Bonelli, Corpus Linguistics at Work, John Benjamins Publishing Amsterdam–Philadelphia, 2001.

[13] Ukrainian web corpus of the University of Leipzig. URL: http://corpora.informatik.uni-leipzig.de/de?corpusId=ukr_mixed_2014

[14] Ukrainian language corpus of the "Chtivo" library. URL: http://korpus.org.ua/

[15] Ukrainian laboratory. URL: https://mova.institute

[16] Corpora of Ukrainian texts. URL: https://lang.org.ua/uk/corpora/

[17] Corpus of the Ukrainian language. URL: http://www.mova.info/corpus.aspx?l1=209

[18] General Regionally Annotated Corpus of Ukrainian (GRAC). URL: uacorpus.org

[19] 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), Volume I: Main Conference, Lviv Ukraine, 2020, pp. 489–506.

[20] V. Starko, Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools, in: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020), Volume I: Main Conference, Lviv Ukraine, 2020, pp. 239–248.

[21] Grammatical Error Correction and Fluency Corpus for the Ukrainian Language. URL: https://github.com/grammarly/ua-gec

[22] The Ukrainian Brown Corpus. URL: https://github.com/brown-uk/corpus

[23] O. Cherednichenko, O. Kanishcheva, O. Yakovleva, D. Arkatov, Collection and Processing of a Medical Corpus in Ukrainian, in: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020), Volume I: Main Conference, Lviv Ukraine, 2020, pp. 272-282. URL: http://ceur-ws.org/Vol-2604/paper21.pdf

[24] N. Grabar, T. Hamon, Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation, in: Proceedings of the 1th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2017), Kharkiv Ukraine. URL: http://colins.in.ua/wp-content/uploads/2017/12/COLINS-2017_conference_proceedings.pdf

[25] N. Grabar, T. Hamon, WikiWars-UA: Ukrainian corpus annotated with temporal expressions, in: Proceedings of the 3d International Conference on Computational Linguistics and Intelligent Systems (COLINS 2019), volume II: Workshop, Kharkiv Ukraine, 2019. URL: http://colins.in.ua/wp-content/uploads/2017/04/preface_volume2_2019.pdf

[26] T. Erjavec, MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora, in: the Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10 ELRA, Paris, 2010.

[27] O. V. Dereza, D. A. Kayutenko, A. S. Fenogenova, Automatic Morphological Analysis for Russian: a Comparative Study, in: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue", 2016. URL: http://www.dialog-21.ru/media/3473/dereza.pdf

[28] A. K. Asiryan, Sravneniye instrumentov morfologicheskoy razmetki [Comparison of tools for morphological marking], Nauchnyy vzglyad v budushcheye. 1.7: Tekhnicheskiye nauki (2017) 27–34. URL: https://doi.org/10.30888/2415-7538.2017-07-01-027 (in Russian)

[29] L. Padró, E. Stanilovsky, FreeLing 3.0: Towards Wider Multilinguality, in: Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA, Istanbul Turkey, 2012. URL: http://nlp.lsi.upc.edu/freeling/node/

[30] M. Weisser, Dialogue Annotation & Research Tool (DART). Version 3.0 (May 2019). URL: http://martinweisser.org

[31] A. Rysin, V. Starko, Large Electronic Dictionary of Ukrainian (VESUM), 2005–2021. URL: https://github.com/brown-uk/dict_uk

[32] AntConc. Build 3.4.4. Laurence Anthony. URL: http://www.laurenceanthony.net/software/antconc/

[33] Sketch Engine. Corpus query and management system. URL: https://www.sketchengine.eu/nosketch-engine/

[34] A. D. Shmelev, Russkiy anekdot: Tekst i rechevoy zhanr: monografiya [Russian anecdote: Text and speech genre: monograph], Moscow, 2002. (in Russian)

[35] R. Charchika, Gender stereotyping, 2020. URL: https://www.researchgate.net/publication/344727936_Gender_Stereotypes. doi: 10.13140/RG.2.2.34843.28969.

[36] V. O. Samokhina, Zhart u suchasnomu komunikatyvnomu prostori Velykoyi Brytaniyi i SSHA: tekstualʹnyy ta dyskursyvnyy aspekty [Joke in the modern communicative space of Great Britain and the United States: textual and discursive aspects] (Germanic languages), Ph.D. thesis, Taras Shevchenko National University of Kyiv, Kyiv, 2010 (in Ukrainian)

[37] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíˇcek, V. Kováˇr, J. Michelfeit, P. Rychlý, V. Suchomel, The Sketch Engine: ten years on, Lexicography, 1(1) (2014) 7–36. doi: 10.1007/s40607-014-0009-9.