DOI: https://doi.org/10.31392/NPU-nc.series9.2022.24.07
UDC: 81'33:37.091.315.7:044.77

Iryna V. Strashko
PhD in Philosophy, Associate Professor,
Head in the Department
of Romance and Germanic Philology
Foreign Philology Faculty,
National Pedagogical Dragomanov University,
Kyiv, Ukraine,
https://orcid.org/ 0000-0001-5137-991X
e-mail: i.v.strashko@npu.edu.ua

MULTIMEDIA CORPUS “EVERYONE HAS THEIR OWN WAR”: CONCEPTION, ANNOTATION, AND PROSPECTS

Bibliographic Description:

Strashko, I. (2022). Multimedia Corpus “Everyone Has Their Own War”: Conception, Annotation, and Prospects. *Scientific Journal of National Pedagogical Dragomanov University. Series 9. Current Trends in Language Development*, 24, 81–93. https://doi.org/10.31392/NPU-nc.series9.2022.24.07

Abstract

The paper focuses on the conception of multimedia corpus “Everyone has their own war”, its marking up with ELAN software, the system of tiers, and types of tasks that can be solved with the help of this corpus. The spoken corpus will contain recordings of semi-directive audio interviews in Ukrainian / Russian presented in audio and text formats, and translations of the recordings into English and French, designed and annotated with ELAN software employment.

The practical result of the work is the creation of an annotation system that explicitly and visually represents the phenomena of speech. Each sample / pattern of original speech is annotated on both lexical and morphological levels. The format of the transcript of oral discourse and a markup by tiers provides the possibility of processing the presented language material. Moreover, at different tiers, there is both purely linguistic information and information of another level – marks about the emotional component, and the selection of non-verbal phenomena.

The next stage of the project aims to correct the existing errors, to improve the system of multi-level annotation and then to integrate all the materials (audio, text and annotation files) into a corpus.

It is expected that as a result the multimedia corpus will be used not only for linguistic research, but also as a source of educational material and a base for both studying and teaching, in particular the Ukrainian language in its sound form, in teaching translation, etc. Since it is assumed that the corpus is *dynamic, the collection of materials continues on a regular basis. The further task of the corpus development is to increase its volume and ensure its gender and age balance. Yet another task is to expand the geography of the corpus by including speech recordings made in different regions of Ukraine. Along with contributing to the development of corpus-based research, corpus creation itself will become a chronicle of modern sociolinguistic stratification of Ukrainian society. Therefore, it can also serve as an informative source for studying the individual experience of the events of the Russian-Ukrainian war.*

Keywords: *multimedia corpus, semi-directive audio interviews, ELAN software, linguistic annotation, oral discourse, Russian-Ukrainian war.*

1. Introduction.

The current state of development of computer technology, its synthesis with linguistics determines its use in speech analysis and an increase in the number of various corpora (multimedia inclusive) in different languages. Despite the undoubted progress made recently by Ukrainian linguists, with the development and availability of several (annotated and non-annotated) text corpora, Ukrainian speech corpora still remain marginalized both in linguistics and in corpus research.

The relevance of the task to compile a collection of personal stories of Ukrainians during the Russian invasion (2022), presented in the form of semi-directive interviews followed by the corpus building, is determined by the lack of a socially representative spoken corpus in Ukraine, with a multidimensional representation of linguistic material. In addition, such a corpus will be of national, cultural, and human significance, as the recorded materials enable keeping in memory the events witnessed and survived by the respondents during the Russian invasion.

The theoretical and methodological basis of the study is formed by the works of numerous researchers among which it is necessary to mention the following: Baude et al. (2006); Bogdanova et al. (2009); Darchuk (2010); Debaisieux (2005); Goedertier et al. (2000); Kibrik & Podlesskaya (2009); Korotaev (2011); Shvedova et al. (2017). The research on semantic annotation in a Ukrainian Corpus (Darchuk et al., 2016; Starko, 2020, 2021) contributed to semantic annotation of collected by us linguistic data. The principles of multilevel annotation used in sound corpora (Asinovsky et al., 2009; Korotaev, 2011; Sherstinova et al., 2009) served a basis for developing the annotation system for the corpus under construction. Particularly useful were the works on self-correction in spoken discourse based on corpus data (Podlesskaya et al., 2019) and the works on corpus instruments in grammatical studies (Lyashevskaya, 2016). Very helpful for work with ELAN annotation tools were the previous studies by foreign and Ukrainian researchers (Crasborn & Sloetjes, 2010; Plahotnikova, 2014; Sloetjes & Wittenburg, 2008), as well as those on the extension of the controlled vocabularies (Crasborn et al., 2012).

In our research we also rely on the studies related to effective methods of recording and analysing interviews especially those dealing with the peculiarities of recording material (Rozental, 2003) and methods of recording and certification of autobiographical narratives (Labashchuk, 2019). Of special interest are studies of various aspects of oral historical interviews, particularly the project “Ukraine of the XX century in the memory of women” by Kys (https://uamoderna.com/images/archiv/11/20_UM_11_Povidomlennia_Kis.pdf) and the research by Hrinchenko who studied narrative autobiographical interviews (2008, 2009).

Plahotnikova’s works (2017, 2018) on the principles of construction and practical implementation of the Corpus of Ukrainian transcribed oral speech have become the creative stimulus for the research.

2. Aim and Objectives.

The paper ***aims*** to present the conception of a multimedia corpus based on the collection of recorded interviews with Ukrainians.

In accordance with the aim of the research, the ***objectives*** are as follows:

- to present the original conception of the spoken corpus;

- to describe the method of collecting linguistic material and the specifics of selecting informants;

- to explain the format of the transcript of oral discourse and describe a markup by tiers;

- to outline the types of tasks that can be solved with the help of the corpus.

3. Methods.

The abovementioned aim determines the employment of the auditory analysis of speech, as well as methods of linguistic annotation (spelling decoding, fixation of extra-linguistic elements of the sound chain, partial and semantic markup, etc.). To enhance the objectivity and reliability of the obtained results the comparative and descriptive methods were used.

4. Results and Discussion.

The multimedia corpus is being developed at the Faculty of Foreign Philology of the National Pedagogical Dragomanov University. The idea of its creation arose at the end of March 2022. At the same time the collection of sound recordings began. The methodology applied to create this spoken corpus combines linguistic, sociological, and historical perspectives. It is based on the model developed by French scientists from the University of Orleans when creating the ESLO corpus (Baude, 2004; Abouda & Baude, 2006).

In total, the interviews with 43 people (18 men and 25 women) aged from 19 to 75 from different regions of Ukraine representing different social and age groups were recorded. All records were made in the genre of thematic audio interviews on the individual perception of the events of the Russian-Ukrainian war, 2022.

Initially, recordings were done in the field with a cell phone or tablet. Later, especially during the long curfew, remotely, we used a laptop with the Internet. Almost 40 hours of audio were recorded, including face-to-face interviews and telephone interviews recorded with the help of Audacity and different messengers (Telegram, Viber, and WhatsApp). All sound recordings were accompanied by metadata. The metadata list is updated as it is filled with data. As part of the metadata were indicated the date and number of recording, an identifier number and personal data of each informant (age, occupation, social status, and place of residence), format and situation of recording. This provides a detailed description of potentially relevant characteristics.

The speech genre of a semi-directive interview is a complex informative genre that contains description of an informant’s life during martial law who recalls and tells about the events, changes in personal life, attitude towards them from the beginning of the full-scale invasion of the Russian Federation into the territory of Ukraine until the moment of communication with the interviewer. A typical interview is structured according to the following general scheme: information about the informant’s place of birth → about how the speaker learned about the outbreak of war → his/her readiness for it → information about the first days of the war → a story about changes in personal life caused by the war (changing the place of residence, the fact to become an internally displaced person / refugee abroad, life in the occupation, etc.) → information about his/her present life → hopes and dreams for the future. This scheme envisages obtaining informants’ answers to the following questions:

Where are you from? Where do you live permanently? How did you find out that war had broken out? Where were you caught by the news of the war? Had you considered the possibility of Russian invasion before it started? Were you prepared for it? How has the war *changed your life? What are you doing now? What are your dreams and hopes for the future?*

Unlike the official interviews, the stories in the collected by us interviews are characterized by more details, i.e. more elaborate events, availability of information not only about informants themselves, but also about their family members, acquaintances, and neighbours.

The oral nature of communication while interviewing causes the transformation of the above outlined scheme, namely: some of its elements may be missing, digressions and stories about related events may be included, comments on a particular event or situation may be added. However, information about the beginning of the war, readiness / unreadiness for the war, dreams for the future are mandatory components of each interview.

In the process of further work on collecting audio material, the original ideas of the project were developed and adjusted. In particular, the geography of recordings was expanded, although initially it had been planned to record only residents of Kyiv and its region. Currently, there are recordings of informants from Kyiv and the region (72%), and also from other regions of Ukraine (28%). In addition, the initially compiled list of questions was also revised: interviews were conducted based on the real-life situations of informants, rather than on the original general scheme.

Since it is assumed that the corpus shall be dynamic, the collection of materials continues on a regular basis. Therefore, one of the tasks is to expand the geography of recordings and ensure gender balance, as the selection of potential project participants is based on territorial, gender, and age criteria.

The recording protocol envisages the preliminary contact with each informant to get acquainted with the purpose of the project, followed by recording the interview. As a prerequisite for recording there was announced the anonymity of all project participants. So, the interaction with each speaker is anonymized by assigning an identifier number associated with the audio recording and all its data. At the initial stage of the project, informants were selected randomly through personal contacts of the interviewer and by recommendations of the informants themselves, as well as their friends, relatives, colleagues or neighbours.

The percentage of men and women as a whole correlates with the current situation in the Ukrainian society caused by the Russian invasion. In particular, the geographical distribution of the population has significantly changed due to the massive internal displacement of people, which still continues. According to available data, the number of women aged 40–69 is 23% higher than of men, while the number of 35–39-year-old men is almost twice as high as of those aged 15–19 (Ekonomichna pravda, 2022 https://www.epravda.com.ua/publications/2022/06/28/688487). In addition, this ratio is largely due to the gender factor as women were more willing to participate in the project. The gender component also affected the duration of the interviews. But for the gender factor, the duration of the interview was also influenced by the informant’s emotional condition, i.e. by the depth of what they had experienced and seen and by the need to speak out and share their feelings. At the end of the interview, the informants were expressing words of gratitude for being listened to, e.g.: “*і вам дякую шо ви мене вислухали// шо ви от робите такі проекти/ які / е-е /надають значення /цій ситуації/ цій війні/ і дають змогу людям які на жаль попали під її вплив/ висказатись/ розказати свою історію ...*” (interview, VUKS_INT_017).

Interviews were recorded in Ukrainian (75%) or Russian language at the informant’s choice. The number of interviews recorded in Russian is currently small (23%, 5 men and 5 women, mostly elderly people, residents of such big cities as Kyiv and Kharkiv). This is due to the fact that after the Russian invasion, informants who speak Russian in everyday life has made an attempt to switch to Ukrainian, and those who are fluent in both languages preferred Ukrainian for recording. There is also one “mixed” interview (2%), partly in Ukrainian, partly in Russian, – the interview given by the woman aged 70, a Bucha resident. It is interesting to note that in response to the question about the prospects for the existence of the Russian language in Ukraine and her attitude toward the Russian language, she said: “у *мене до української мови добре ставлення // ви до мене звернулись на російській /я вам відповіла// зараз я знаходжусь в-в Германії / і тут ну коли зустрічаєш українців / ми / ну я не знаю /якось так виходить шо ми переходимо на українську мову// ми не = ми спілкуємося тільки на українській мові // тут і з Енергодару у мене жіночка з-з дітьми /знайомі... вже так роззнайомилися... ми всі говоримо українською мовою// і коли до нас звертаються німці через перекладач / вони до нас звертаються / вони пишуть нам / ну за =питання/ і в перекладачі на українській мові відповідь// я думаю шо є// противно навіть говорити цією мовою/ але ж знаєте /звичка// коли вчився / і в школі /і все життя / говорили ми ... на ці ... ви ж пам'ятаєте як раніше було// як говориш українською мовою то ти жлоб/// говорили всі на російській мові /а зараз я... я вважаю що наша мова сама найкраща в світі // і взагалі Україна сама найкраща країна на всьому білому світі //”(*interview, VUKS_INT_024).

The above cited excerpt is the “raw” transcript of a fragment taken from the recording to be fixed at Speaker Phrase tier.

It is worth mentioning that the use of a phone or tablet at the initial stage often led to the fact that the quality of the sound material was not uniform. In addition, a certain percentage of recordings of low quality is the result of unstable Internet connection.

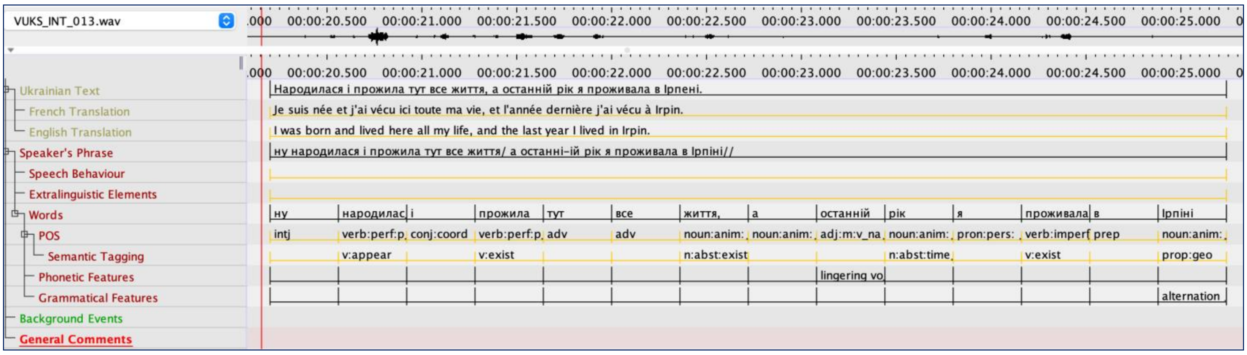
All telephone recordings are saved in the format .wav, with a sampling rate of 44100 Hz and a 16-bit resolution.

The primary processing of the speech material involved the elimination of noise fragments that did not contain speech, adjusting the quality of the recordings, their decoding and transforming into text form.

All recordings were decrypted in two stages. At the first stage it was done with the help of Transcriptor, an online speech-to-text converter (Transcriptor, 2022 https://transkriptor.com). At the second stage, the obtained texts were manually edited / corrected by the author of the project assisted by the students of the Department of Applied Language Studies, Comparative Linguistics, and Translation.

Given the specificity of the empirical object, like many other researchers (Asinovsky et al., 2009; Kibrik & Podlesskaya, 2009; Korotaev, 2011; Plahotnikova, 2017; Sherstinova et al., 2009), the author of the project has faced the problem of its adequate representation for linguistic analysis on a systematic basis.

The basic principle of creating a set of tiers is a comprehensive, visual representation of each sample of oral speech, given all the information that can be required for analysis. Linguistic annotation is carried out with ELAN software (2022 https://archive.mpi.nl/tla/elan). This software is designed for multi-level annotation and therefore offers flexible granularity and heterogeneous linguistic material representation. The desire for an integral representation has led to the epistemological orientation of tiers, among which are as follows: Ukrainian Text, English Translation, French Translation, Speaker’s Phrase, Words, POS, Semantic Tagging, Phonetic Features, Grammatical Features, Speech Behaviour, Extralinguistic Elements, Background Events, and General Comments.



Ukrainian Text: Народилася і прожила тут все життя, а останній рік я проживала в Ірпені.

French Translation: Je suis née et j'ai vécu ici toute ma vie, et l'année dernière j'ai vécu à Irpin.

English Translation: I was born and lived here all my life, and the last year I lived in Irpin.

Speaker's Phrase: ну народилася і прожила тут все життя/ а останні-ій рік я проживала в Ірпіні//

Speech Behaviour

Extralinguistic Elements

Words	ну	народилас	і	прожила	тут	все	життя,	а	останній	рік	я	проживала	в	Ірпіні
POS	intj	verb:perf:p.	conj:coord	verb:perf:p.	adv	adv	noun:anim:	noun:anim:	adj:m:v_na	noun:anim:	pron:pers:	verb:imperf	prep	noun:anim:
Semantic Tagging		v:appear		v:exist			n:abst:exist			n:abst:time		v:exist		prop:geo
Phonetic Features									lingering vo					
Grammatical Features														alternation

Background Events

General Comments

Figure 1. The example of multi-level markup in ELAN

Linguistic annotations are filled in as a relevant phenomenon is fixed. Annotation is carried out both manually and automatically (in case of tokenizing).

The practice of creating corpora of spoken language (Kibrik & Podlesskaya, 2009) shows that transcription of spoken language is a very laborious and time-consuming process, that is why the main way of representing spoken language is orthographic spelling with the reflection of its individual features (to be discussed below), without marking up the prosodic features of speech. The fact that the texts will be given in standard spelling, without the use of any transcription system, makes them accessible to any user. This removes the problem of developing a transcription system for a spoken text.

At the initial stage the segmentation of the spoken text was supposed to be carried out for the purpose of readability. Therefore, the initial unit of segmentation was a syntactic unit corresponding to a traditional sentence. Those ones were obtained as a result of auditory analysis of previously decrypted spoken texts. Punctuation marks were placed based on both the subjective perception of the boundaries of the utterance as a complete segment of speech (the conventional and usual marker in the traditional sense of which is a pause as a suprasegmental phenomenon) and on the rate of speech of a concrete speaker. It is worth mentioning here the experience of speech segmentation during the creation of the Corpus of Ukrainian transcribed oral speech. The authors of this corpus faced with the difficulties in correct and consistent displaying of annotations due to speech division into sequences of phonetic units (See: Plahotnikova, 2017, p. 118).

In contrast to the approach used in most of spoken corpora (Sherstinova et al., 2009), the oral texts were not divided into fragments, as it would destroy the very idea of the interview as a conceptually unified structure in which "[…] the experience of a separate event is always embedded in a context, a biographical construction, united by a common meaning" (Rozental, 2003, p. 326). Moreover, the existence of tiers designed to translate a spoken text into English and French makes it necessary to keep segmentation at the sentence level, not to mention the fact that the corpus materials are supposed to be used for educational purposes, namely in translation studies. In order to save this possibility, three tiers are interconnected, with translation tiers being dependent on Ukrainian Text, which is the parent tier. This constrains presenting the translated text versions in English and French close to the literary version of the Ukrainian language.

The preliminary analysis of the material has shown that segmentation and the punctuation of the spoken text is problematic and can be implemented in different ways. For example, the option of segmenting the spoken text, despite its predominantly monological nature, proposed by a student (1) who participated in the project, differed from the option when the boundaries of the utterance were determined by the author of this research (2), who was well acquainted with the original:

1) до речі/ я справді боялася почути їх все життя/ тому що я пам'ятала ще розповіді бабусі та дідуся// вони пережили Другу світову // і бабуся завжди казала… головне/ щоб не було війни/ і оці мамині слова / зранку коли війна почалася/ я... просто / ну це був шок/ тотальний шок...

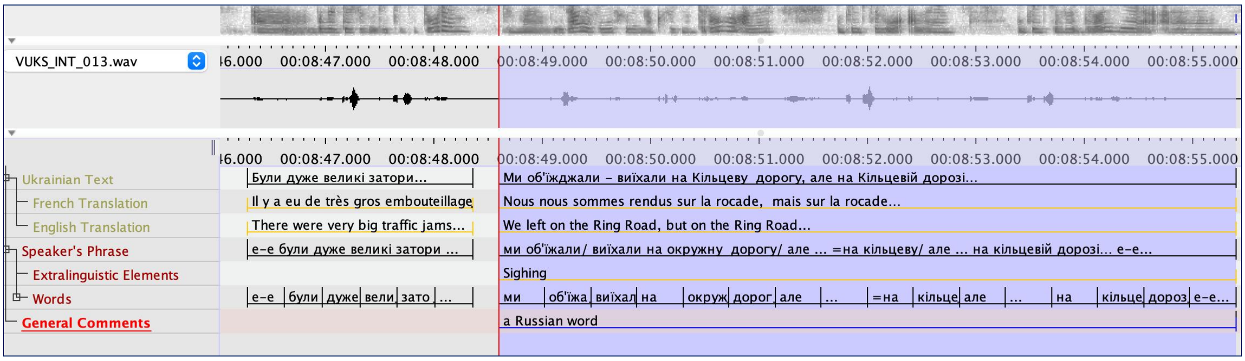
2) до речі/ я .../ справді я боялася почути їх все життя/ тому що-о я пам'ятала ще розповіді бабусі та дідуся// вони пережили Другу світову/ і бабуся завжди казала / головне щоб не було війни// і-і оці мамині слова зранку/ коли во = війна почалася / я…просто ...ну це був шок/ тотальний шок... (interview, VUKS_INT_013).

Different vision of the place of the utterance boundary may be explained by the fact that during the segmentation a word or a phrase can be related to both the preceding and the following context.

Since the creation of only written record of spoken language "inevitably leads to the substitution of sound substance, oriented to auditory perception, with its graphic record, which is decoded visually" (Ratnikova, 2012, p. 4), it has become necessary to create a new tier (Speaker Phrase), which would record the features of a specific speech activity of the speaker, in particular the simultaneous building of his thought and its verbalization.

The Speaker Phrase tier displays exactly the peculiarities that were uttered by the speaker. If in reality the sound is not realized, it is marked only at the Text level. The creation of this tier increases the possibility of analysing spontaneous speech and its comparison with the codified literary language.

This tier makes it possible to indicate the shortcomings of oral speech, deviations from the codified literary norm that are perceived by ear at the phonetic level (*шо, тіпа, всьо, капєц, Сірьожа,* etc.), language phenomena at the lexical level, especially russianisms (*вооруженія, предпосилки, замужем, часний, чувствовалось, одіялах*), signs of spontaneous speech, graphic marks of hesitation pauses (*гм, е-е, ну*) and speech failures (namely self-correction), and remarks describing the process of speech generation. Punctuation marks are not used at this level of representation.



Ukrainian Text	Були дуже великі затори…	Ми об'їжджали – виїхали на Кільцеву дорогу, але на Кільцевій дорозі…
French Translation	Il y a eu de très gros embouteillage	Nous nous sommes rendus sur la rocade, mais sur la rocade…
English Translation	There were very big traffic jams…	We left on the Ring Road, but on the Ring Road…
Speaker's Phrase	е-е були дуже великі затори …	ми об'їжали/ виїхали на окружну дорогу/ але … =на кільцеву/ але … на кільцевій дорозі… е-е…
Extralinguistic Elements		Sighing
Words	е-е були дуже вели зато …	ми об'їжа виїхал на окруж дорог але … =на кільце але … на кільце дороз е-е…
General Comments		a Russian word

Figure 2. A markup example at seven annotation levels

The above example shows the difference in the presentation of the same oral text at several annotation levels, in particular close to literary text variant in the Ukrainian language, which is an independent type, its translation into French and English (both being of Dependent Type), as well as some peculiarities of the informant's utterance represented at the Speaker's Phrase tier (depending on Ukrainian Text), Extralinguistic Elements, Words, and General Comments.

This example also serves as a case of self-correction, when the speaker, after using a Russian word (*окружну*), understood the mistake and corrected it by herself (the case of self-correction is marked by the symbol =). This word is also marked as "*a Russian word*" at

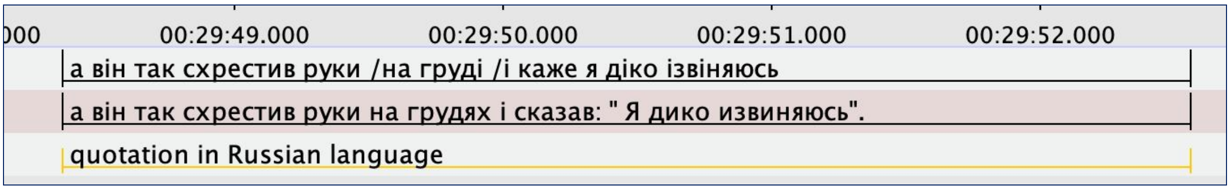
General Comments tier. As shown in the example, the chosen way of presenting the peculiarities of speaker's speech includes several symbols. For instance, the incompleteness of the phrase is marked by a dotted line "..."; speech segmentation into phrases and syntagms is shown in traditional way: a single slash "/" is a sign of division into syntagms and a double slash "//" means the end of the phrase. Also, in this example with the symbol "*e-e*" the filled hesitation pauses are marked. As we can see the proper names are capitalized.

The General Comments tier (an Independent type) contains other information that was not reflected at other levels, or was reflected partially, but is important for finding out the reasons for possible gaps in the annotations. For instance, here unrecognized fragments of spoken discourse (word, phrase or whole sentence) are marked with the remark "unintelligible". If the unintelligible fragment is longer than a word, the start and end time need to be indicated.

Two more tiers are designed to mark the emotional condition of the speaker and various non-speech phenomena in real communication circumstances to reflect the naturalness of speech. At the tier of Extralinguistic Elements (see the above example), non-verbal events such as sighing, laughing, coughing, yawning, smacking, etc. are fixed. Based on the observations, we can say that these phenomena are caused by the speaker's emotional feelings overwhelming them during the act of speech communication.

Words tier (depends on Speaker's Phrase) provides automatic segmentation into words defined by ELAN software parameters.

The Speech Behaviour tier is aimed at reflecting the speaker's speech behaviour during the interview, which helps to understand the semantic connotations of the content of the utterance. Also at this level, citations are displayed, including those in other languages (mainly Russian or English). For instance, the example below displays the fragment of the informant's utterance who cites the words of another person in the Russian language and imitates a Chechen and Russian accent with deliberately distorting the words. Thus she tries to make the situation more real.



а він так схрестив руки /на груді /і каже я діко ізвіняюсь

а він так схрестив руки на грудях і сказав: " Я дико извиняюсь".

quotation in Russian language

Figure 3. A markup example for an informant's speech behaviour

At the Background Events level we provide comments on the communication situation and events that accompanied the recording (birds singing, dogs barking when recording in the park, or the sound of an explosion near the interviewer's house when recording over the internet).

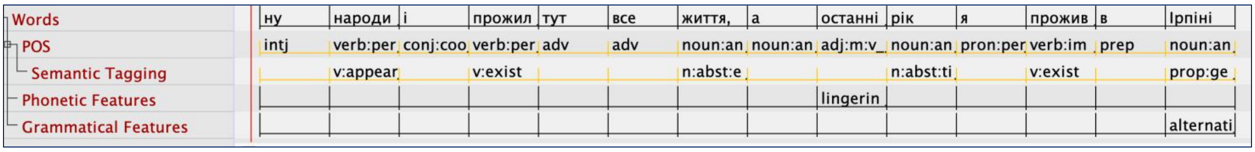
The Phonetic Features tier (depending on Words) marks the peculiarities of phonetic realization already in the primary data markup. Here, abnormal accents or accents in words with possible variants (syllables, compositions), chanting are noted. In case of incorrect stress or possible variations, the vowel is highlighted with a capital letter. Chanting and lingering vowels (see example below, fig. 4) and consonants are conveyed with hyphens: *останні-ій*.

The Grammatical Features tier (depending on Words) is necessary for fixing errors in the grammatical structure of the utterance. The word forms used grammatically incorrectly (wrong case, gender, number), which have been transformed into correct ones at the Ukrainian Text tier, are registered here. But since the tagging system in Controlled

Vocabulary does not provide for the marking of such “real” sounding forms, without the introduction of this level such forms would remain without any marking at all, which would significantly impoverish the prospects for studying oral speech, in particular, in correlation with undistorted, dictionary ones.

Preliminary analysis of the collected material indicates the presence of interference errors of phonetic and grammatical nature in informants’ speech. Moreover, grammatical errors are often associated with phonetic and lexical deviations.

At the tier of part-of-speech (POS) markup, which depends on Words (Symbolic Association), each word is assigned a set of tags. For morphological annotation, the tagging system developed for the Large Electronic Dictionary of Ukrainian (https://r2u.org.ua/vesum) was taken as a basis (almost without changes, with some exceptions). For this purpose, special Controlled Vocabularies (CV) were defined. Like any Controlled Vocabulary, all CVs are associated with a “linguistic type” specification for two tiers: the tier of POS tagging and that of Semantic Tagging. Similarly to other projects (Sloetjes & Wittenburg, 2008), each element in each CV is stored as an attribute of the annotation referring to its element. The value of the element is stored in the format .eaf so that it is available for visualization and search.



Words	ну	народи	і	прожил	тут	все	життя,	а	останні	рік	я	прожив	в	Ірпіні
POS	intj	verb:per	conj:coo	verb:per	adv	adv	noun:an	noun:an	adj:m:v_	noun:an	pron:per	verb:im	prep	noun:an
Semantic Tagging		v:appear		v:exist			n:abst:e			n:abst:ti		v:exist		prop:ge
Phonetic Features									lingerin					
Grammatical Features														alternati

Figure 4. A markup example at five annotation levels
(words, POS, semantic tagging, phonetic and grammatical features)

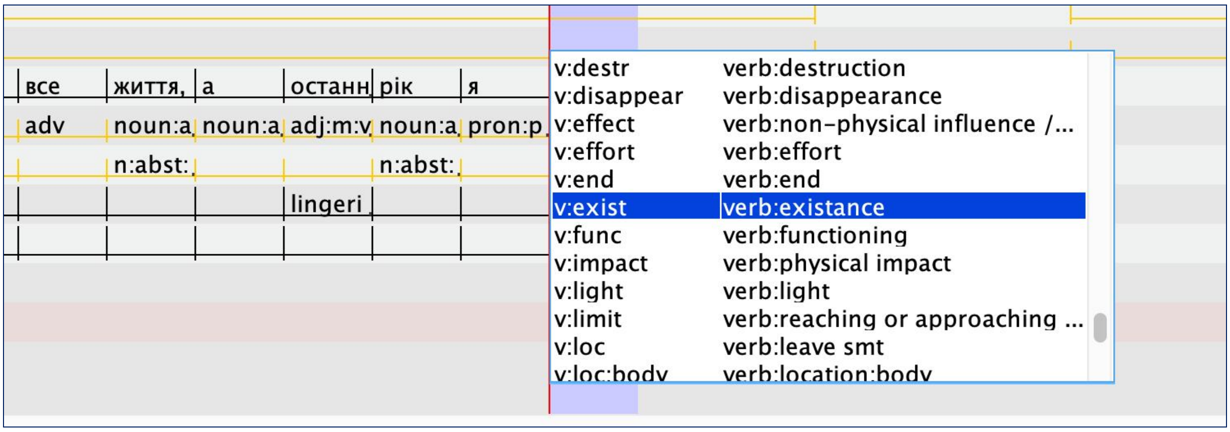
It is also worth mentioning that CVs for part-of-speech tagging for both English and French languages were defined as well, but the necessity to do such markup still remains open.

As in the semantic markup of the General regionally annotated corpus of the Ukrainian language (http://uacorpus.org/Kyiv/ua/rozmitka-tekstiv/semantichna-anotaciya), semantic tags are developed separately for six large word categories: concrete nouns, abstract nouns, proper names, adjectives, adverbs, and verbs.

In accordance with the classification adopted in GRAC, semantic tags within each group are assigned in the following order:

concrete nouns (conc) – taxonomy, mereology, topology, evaluation;
abstract nouns (abst) – taxonomy, mereology, evaluation;
proper names (prop) – taxonomy;
adjectives – taxonomy, assessment;
adverbs – taxonomy, evaluation;
verbs – taxonomy, causation (ibid.).

At the tier of semantic markup, which is dependent on POS by Symbolic Association type, each word is assigned a categorized tag (see an example below):



все | життя, | а | останн | рік | я

adv | noun:a | noun:a | adj:m:v | noun:a | pron:p

n:abst: | n:abst:

lingeri

v:destr	verb:destruction
v:disappear	verb:disappearance
v:effect	verb:non-physical influence /...
v:effort	verb:effort
v:end	verb:end
v:exist	verb:existance
v:func	verb:functioning
v:impact	verb:physical impact
v:light	verb:light
v:limit	verb:reaching or approaching ...
v:loc	verb:leave smt
v:loc:body	verb:location:body

Figure 5. An example of a Controlled Vocabulary for semantic markup

It should be emphasized that the system of linguistic annotation within the ELAN software is being tested and refined on a continuous basis. Already to date our experience in speech segmenting shows that in perspective it will be advisable to introduce a tier with the informant’s and the interviewer’s speech overlapping. There is a need for latter when the second speaker has started an utterance while the first one is still talking.

As in annotation files of ELAN within the developed linguistic annotations, the search options are possible within one or more annotation files located in the same directory, the user has an opportunity to search for a word, for a semantic or a morphological tag.

5. Conclusion.

The multimedia corpus, which is currently being built, primarily aims at analysing the morphological and semantic phenomena related to the discoursive characteristics of informants’ speech. It will contain recordings of audio interviews in Ukrainian / Russian, presented in audio and text formats, translated into English and French, designed and annotated with ELAN software.

The main theoretical and technical decisions made in the corpus (transcripts, linguistic annotation) correspond to specific research and educational purposes. The practical result of the work is the creation of the annotation system that explicitly and visually represents the phenomena of speech. Each example of speech is annotated at ~~on~~ lexical and morphological levels, as well as at the level of translation. The format of the transcript of oral discourse and the markup by tiers provides the possibility of processing the presented language material. Moreover, at different tiers, there is both purely linguistic information and extralinguistic information – marks about the emotional component, and the selection of non-verbal phenomena: laughter, coughing, filled and unfilled hesitation pauses, etc.

It is expected that the built multimedia corpus can be used not only for linguistic research, but also as a source of educational material and a base for both studying and teaching, in particular the Ukrainian language in its sound form, in teaching translation, etc. At the next stage of the project, the correction of the existing errors and improvement of the system of multi-level marking is foreseen.

Along with contributing to the development of corpus-based research, corpus creation itself is part of the social and oral history of the Ukrainian society. Therefore, the created corpus will also serve as an informative source for studying the individual experience of the events of the Russian-Ukrainian war.

The further task of the corpus development is to increase its volume and ensure its gender and age balance. Yet another task is to expand the geography of the corpus by including speech recordings made in different regions of Ukraine.

References

Abouda, L., & Baude, O. (2006). Constituer et exploiter un grand corpus oral: choix et enjeux théoriques. Le cas des ESLO. [Building and using a large oral corpus: choices and theoretical issues. The case of ESLOs]. *Corpus en Lettres et Sciences sociales, Des documents numériques à l'interprétation.* Albi, France. halshs-01162506 [in French].

Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., & Sherstinova, T. (2009). The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation. *International Conference on Text, Speech and Dialogue* (September, 2009, Springer, Berlin, Heidelberg), 250–257.

Baude, O. (2004). Les corpus oraux entre science et patrimoine. L'expérience de l'Observatoire des pratiques linguistiques. [Oral corpora between science and heritage. The experience of the Observatory of Linguistic Practices]. *Publicisation de la science*. Grenoble, France. halshs-01162520. [in French].

Baude, O., Blanche-Benveniste, C., Calas, M. F., Cappeau, P., Cordereix, P., Goury, L. & Mondada, L. (2006). Corpus oraux, guide des bonnes pratiques. [Oral corpus, guide to good practice]. 203 p. CNRS Editions, Presses Universitaires Orléans. [in French].

Bogdanova, N. V., Asinovskiy, A. S., Rusakova, M. V., Ryko, A. I., Stepanova, S. B., & Sherstinova, T. Yu. (2009). Zvukovoy korpus kak sposob monitoringa i fiksatsii raznykh form yestestvennogo yazyka. [Sound corpus as a way to monitor and record different forms of natural language]. *Komp'yuternaya lingvistika i intellektual'nyye tekhnologii*, (8), 15. [in Russian].

Crasborn, O. A. & Sloetjes, H. (2010). Using ELAN for annotating sign language corpora in a team setting. In: *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages:Corpora and Sign Language Technologies*, LREC 2010, 22-23 May 2010, Malta.

Crasborn, O., Hulsbosch, M., & Sloetjes, H. (2012). Linking Corpus NGT annotations to a lexical database using open source tools ELAN and LEXUS. *Sign-lang@ LREC 2012*). *European Language Resources Association* (ELRA), 19–22.

Darchuk, N. (2010). 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*, 21, 45–49. [in Ukrainian].

Darchuk, N., Zuban', O., Lanhenbakh, M., & Khodakivs'ka, YA. (2016). AHAT-semantyka: semantychne rozmichuvannya Korpusu ukrayins'koyi movy. [AGAT-semantics: semantic marking of the Corpus of the Ukrainian language]. *Ukrayins'ke movoznavstvo*, 1, 92–102. [in Ukrainian].

Debaisieux, J. M. (2005). Les corpus oraux: situation, exploitation linguistique, bilan et perspectives. [The oral corpora: situation, linguistic exploitation, assessment and perspectives]. *Scolia*, (19), 9–40. [in French].

Ekonomichna pravda. Retrieved June, 30, 2022 from https://www.epravda.com.ua/publications/2022/06/28/688487.

ELAN (Version 6.4) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved June 29, 2022, from https://archive.mpi.nl/tla/elan.

Goedertier, W., Goddijn, S. M., & Martens, J. P. (2000). Orthographic Transcription of the Spoken Dutch Corpus. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece. European Language Resources Association (ELRA).

Heneralnyi rehionalno anotovanyi korpus ukrainskoi movy (HRAK) / M. Shvedova, R. fon Valdenfels, S. Yaryhin, M. Kruk, A. Rysin, V. Starko, M. Vozniak. Kyiv, Oslo, Yena, 2017–2019. [General regionally annotated corpus of the Ukrainian language (GRAC)]. Retrieved August, 12, 2022, from http://uacorpus.org/Kyiv/ua.

Hrinchenko, H. (2008). (Avto) biohrafichne interviu v usnoistorychnykh doslidzhenniakh: do pytannia pro teoriiu naratyvnoho analizu [(Auto) biographical interview in oral history research: to the question of the theory of narrative analysis]. *Skhid-Zakhid: Istoryko-kulturolohichnyi zbirnyk*, 11–12, 59–76. [in Ukrainian].

Hrinchenko, H. (2009). Avtobiohrafichni konstruktsii ta stratehii samoreprezentatsii v interviu-spohadakh kolyshnikh ukrainskykh ostarbaiteriv (poperedni rezultaty doslidzhennia) [Autobiographical Constructions and Strategies of Self-Representation in the Interview-Memoirs of Former Ukrainian Ostarbeiters (Preliminary Results of the Study)]. *Storinky voyennoyi istoriyi Ukrayiny*: Zb. nauk. st., 12, 65–72. [in Ukrainian].

Kibrik, A. & Podlesskaya, V. (Eds.). (2009). Rasskazy o snovideniyakh: Korpusnoye issledovaniye ustnogo russkogo diskursa. [Dream Stories: A Corpus Study of Oral Russian Discourse]. Litres Editions. 736 s. [in Russian].

Korotayev, N. A. (2011). “Rasskazy o snovideniyakh”: ot avtonomnoy transkriptsii k mnogourovnevoy diskursivnoy razmetke. [“Dream Stories”: from autonomous transcription to multilevel discursive markup]. *Korpusnaya lingvistika*, 205–210. [in Russian].

Labashchuk, O. (2019). Perspektyvy zastosuvannya metodyky naratyvnoho interv’yu dlya suchasnykh fol’klorystychnykh doslidzhen’ [Prospects for using the narrative interview technique for modern folkloristic research]. *Studia Methodologica*, 48, 41–50. [in Ukrainian].

Lyashevskaya, O. (2016). *Korpusnyye instrumenty v grammaticheskikh issledovaniyakh russkogo yazyka* [Corpus instruments in grammatical studies of the Russian language]. Litres Editions. 520 s. [in Russian].

Plakhotnikova, O. (2014). Vykorystannia prohramy Elan v roboti zi zvukozapysamy korpusu ukrainskoho usnoho movlennia [The use of the Elan program in working with sound recordings of the corpus of Ukrainian oral speech]. *Ukrainske movoznavstvo*, 44, 238–243. [in Ukrainian].

Plakhotnikova, O. Yu. (2017). Korpus ukrainskoho transkrybovanoho usnoho movlennia: zasady stvorennia [Corpus of Ukrainian transcribed oral speech: principles of creation]. *Naukovyi visnyk Drohobytskoho derzhavnoho pedahohichnoho universytetu imeni Ivana Franka. Seriia: Filolohichni nauky (movoznavstvo)*, 7, 145–148. [in Ukrainian].

Plakhotnikova, O. Yu. (2018). *Korpus ukrainskoho usnoho movlennia: teoretychni zasady pobudovy y osnovy praktychnoho vtilennia* [Corpus of Ukrainian oral speech: theoretical principles of construction and bases of practical implementation]: dys. ... kand. filol. nauk: 10.02.01 / Plakhotnikova Olena Yuriivna; Kyiv. nats. un-t im. Tarasa Shevchenka. Kyiv, 280 s. [in Ukrainian].

Podlesskaya, V. I., Korotayev, N. A., & Mazurina, S. I. (2019). Samoispravleniya govoryashchego v russkom monologicheskom i dialogicheskom diskurse: opyt korpusnogo issledovaniya [Self-correction of the speaker in Russian monologue and dialogic discourse: the experience of corpus research]. *Komp’yuternaya lingvistika i intellektual’nyye tekhnologii*, 547–561. [in Russian].

Ratnikova, Ye I. (2012). *Prosodicheskaya organizatsiya frantsuzskoy ustnoporozhdayemoy rechi (eksperimental’no-foneticheskoye issledovaniye na materiale radio-i teleinterv’yu)* [Prosodic organization of French oral speech (an experimental phonetic study based on radio and television interviews)] [in Russian].

Rysin, A. & Starko, V. (2005–2022). *Velykyy elektronnyy slovnyk ukrayins'koyi movy*. Veb Versiya 5.9.2. 2005-2022 [Large Electronic Dictionary of Ukrainian]. Web version 5.9.2. Retrieved from VESUM, Large Electronic Dictionary of Ukrainian website, https://r2u.org.ua/vesum/

Rozental’, G. (2003). Rekonstruktsiya rasskazov o zhizni: printsipy otbora, kotorymi rukovodstvuyutsya rasskazchiki v biograficheskikh narrativnykh interv’yu [Reconstruction of life stories: selection principles that guide storytellers in biographical narrative interviews]. *Khrestomatiya po ustnoy istorii*. SPb.: YEU, 322–356. [in Russian].

Sherstinova, T. Yu., Ryko, A. I., & Stepanova, S. B. (2009). Sistema annotirovaniya v zvukovom korpuse russkogo yazyka “Odin rechevoy den” [Annotation system in the sound corpus of the Russian language “One speech day”]. *Formal’nyye metody analiza russkoy rechi. Materialy XXXVIII Mezhdunarodnaya filologicheskaya konferentsiya.* SpbGU, 66–75. [in Russian].

Sloetjes, H. & Wittenburg, P. (2008). Annotation by category-ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).*

Starko, V. (2020). Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools. *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020).* Volume I: Main Conference. Lviv, Ukraine, April 23-24, 2020.

Starko, V. (2021). Implementing Semantic Annotation in a Ukrainian Corpus. CEUR Workshop Proceedings. *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021).* Volume I: Main Conference. Kharkiv, Ukraine, April 22-23, 2021, 435–447.

Transcriptor, 2022. Retrieved April 07, 2022, from https://transkriptor.com.

Vidnovliuiuchy vlasnu pamiat: proekt “Ukraina XX stolittia u pamiati zhinok” Mizhnarodnyi proekt “Zhinocha pamiat” [Restoring our own memory: the project “Ukraine of the 20th century in the memory of women” International project “Women’s Memory”]. Retrieved August 07, 2022, from https://uamoderna.com/images/archiv/11/20_UM_11_Povidomlennia_Kis.pdf [in Ukrainian].

Sources of Illustrative Material

Interview, VUKS_INT_013.
Interview, VUKS_INT_017.
Interview, VUKS_INT_024.

Бібліографічний опис:

Страшко, І. В. (2022). Мультимедійний корпус “Війна у кожного своя”: концепція, анотування, перспективи. *Науковий часопис Національного педагогічного університету імені М. П. Драгоманова. Серія 9. Сучасні тенденції розвитку мов*, 24, 81–93. https://doi.org/10.31392/NPU-nc.series9.2022.24.07.

Анотація

У статті висвітлено концепцію мультимедійного корпусу “Війна у кожного своя”, особливості його розмітки з використанням програмного забезпечення ELAN, систему рівнів, типи завдань, які можуть бути вирішені за допомогою цього корпусу. Корпус міститиме записи напівдирективних аудіоінтерв’ю українською / російською мовами, представлені в аудіо- та текстовому форматах, перекладені англійською та французькою мовами, оформлені та анотовані за допомогою програмного забезпечення ELAN.

Практичним результатом роботи є створення системи анотування, яка експліцитно репрезентує явища мовлення. Кожний фрагмент мовлення мовами оригіналу анотується на лексичному та морфологічному рівнях. Формат стенограми усного мовлення та розмітка за рівнями забезпечують можливість опрацювання мовного матеріалу. На різних рівнях представлено як суто лінгвістичну інформацію, так і позначки про емоційну складову мовця, виділено невербальні маркери.

Наступний етап проєкту спрямований на виправлення наявних помилок, вдосконалення системи різнорівневої розмітки й об’єднання матеріалів (звукових, текстових і анотаційних файлів) у корпус.

Очікується, що створюваний мультимедійний корпус, може бути використаний для лінгвістичних досліджень, навчання перекладу та як джерело навчального матеріалу для вивчення фоно-просодичного рівня української мови і мовлення. Оскільки передбачається, що корпус буде динамічним, збір матеріалу триває. Тому подальшим завданням проєкту є збільшення обсягу корпусу, забезпечення гендерної та вікової збалансованості інформантів і розширення географії корпусу за рахунок включення записів, зроблених у різних регіонах України. Створення корпусу не лише сприятиме розвитку корпусних досліджень, а й поставатиме літописом сучасної соціолінгвістичної стратифікації українського суспільства. Отже, цей корпус також слугує інформативним джерелом для вивчення індивідуального досвіду переживання подій російсько-української війни.

Ключові слова*: мультимедійний корпус, напівдирективні аудіоінтерв’ю, програмне забезпечення ELAN, лінгвістичне анотування, усний дискурс, російсько-українська війна.*