Computer linguistic system modelling for Ukrainian language processing

Victoria Vysotska

Lviv Polytechnic National University, Stepan Bandera 12, 79013 Lviv, Ukraine

Abstract

The general structure of the computer linguistic system (CLS) processing of textual content in the Ukrainian language and the conceptual scheme/model of the functioning of a typical CLS based on modelling the interaction of the main processes and IS components have been developed. Modelling of the main NLP processes of CLS was carried out due to the interaction of the main processes/components of IS and methods of linguistic processing of text content adapted to the Ukrainian language based on grapheme, morphological, lexical, syntactic, semantic, structural, ontological and pragmatic analysis, which allowed to improve the IT of intellectual analysis of the text flow for solving a specific NLP problem. This ensured the adaptation of NLP processes for the analysis of Ukrainian-language textual content. A formal model of a computer linguistic system for processing Ukrainian-language textual content was developed and described, which made it possible to determine the main structural elements and operators of natural language processing at each level of text analysis such as grapheme/phonological, morphological, syntactic, semantic, referential, structural, ontological and pragmatic. Due to the complexity of the morphology of the Ukrainian language, detailed attention is paid to the description of the model of morphological analysis of textual content. Examples of modelling processes for solving typical NLP problems such as CLS identification of viral news headlines and correction of grammatical and stylistic errors are given.

Keywords

computer linguistics, system, NLP, Ukrainian language, information resource, system modelling

1. Introduction

Computer linguistic system (CLS) based on NLP methods for text/audio data analysis is already an integral part of human everyday life [1-5]. On behalf of the user, some CLSs browse the large volume of Internet information and offer new personalized mechanisms/techniques/tools for interacting with the computer [6-9], for example through spam filters of e-mail traffic [10-15], IISS, virtual personal assistants, automatic translation IS etc. CLS with the support of natural language analysis is at the intersection of experimental research [16-21] and practical development of usually commercial software [22-27]. CLS of speech analysis and text analytics interact directly with the user through the

CLW-2024: Computational Linguistics Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2024), April 12–13, 2024, Lviv, Ukraine

victoria.a.vysotska@lpnu.ua (V. Vysotska)

0000-0001-6417-3689 (V. Vysotska)

support of feedback, which significantly and continuously affects the functioning of the software and the results of the analysis [28-33].

The potential for implementing natural language analysis in CLS/IS/modules is constantly growing exponentially [34-38]. A disproportionately large volume of NLP applications is usually implemented by large campaigns due to the complexity of the projects and the need for their commercialization [39-45]. As the opportunities for implementation in everyday CLS become more widespread, they become less visible, masking the complexity of their implementation. In parallel, the development of big data science and computer linguistics, especially based on non-English natural text corpora, has not yet reached the level necessary for simplification, optimization, and standardization of the processes of developing appropriate linguistic software [46-51].

CLSs for solving a large volume of specific NLP problems are just starting to spread and will eventually automate more processes that are currently solved through additional forms and selecting/clicking options/buttons. To develop the IT implementation of the appropriate linguistic software and ensure high reliability of CLS, it is necessary to take into account modern promising scientific methods of ML, data analysis, big data, and based on hypothesis analysis [52-55].

2. Related works

To support the functioning of a typical CLS and the operation of the main processes when solving a specific NLP problem, it is necessary and sufficient to implement the main subsystems as client, server and technological (Fig. 1) [56-63].

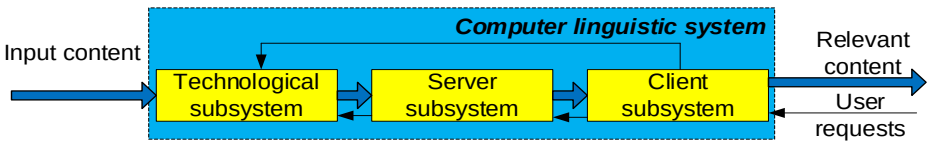

Figure 1: Conceptual diagram of the functioning of a typical CLS

The main processes of the functioning of a typical CLS based on the intellectual analysis of a text stream for the solution of a specific NLP problem [64-72]:

- technological processing of incoming content streams:
    a. search and recognition of content from relevant sources;
    b. accumulation of analyzed content from the source in the cloud;
    c. saving information about the location of the found content in the corresponding source in the database/datastore;
    d. preliminary processing of recognized content in the cloud;
    e. analysis and marking/classification of recognized content according to the degree of relevance to the content and purpose of the CLS;
    f. integration of content provided that its degree of relevance/relevance is greater than the threshold value;
    g. saving integrated relevant content in the DB;

h. forming an image (descriptive service data) of integrated relevant content and saving it in the DB;

- content management based on text analysis and processing through the server subsystem (Fig. 2) based on data from the client subsystem and the content support module:
    a. processing of user request streams from the client subsystem to form the correct IIS expression and subsequent caching of popular content through the server subsystem;

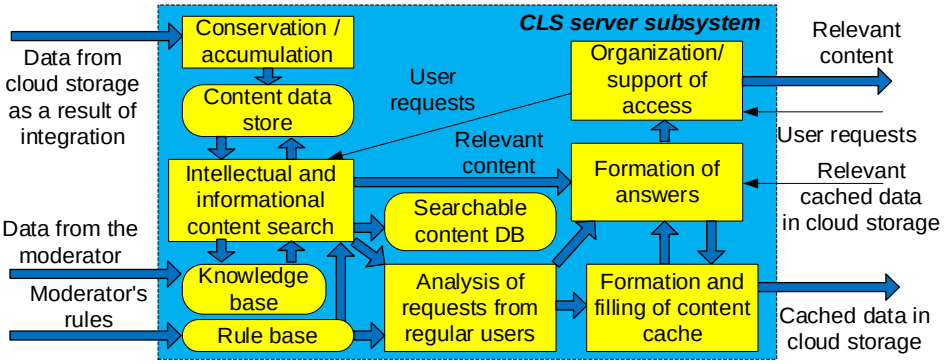

Figure 2: The general conceptual diagram of the functioning of the server subsystem

b. generation of a set of operational relevant up-to-date reports according to the relevant requests of CLS/Website users;
c. analysis of popular user requests at certain time intervals to generate standard reports cached in the cloud;
d. support of effective IIS-relevant content according to user/visitor CLS/Website requests, in particular, IIS implementation of lexical, symbolic, attributive, associative, linguistic, etc.;
e. managing the interaction of CLS content and resource elements;
f. cloud computing support for prompt access to CLS/Website repositories/databases/rules/filters;

- content support based on the analysis and synthesis of information, including service content for the solution of a specific NLP problem:
    a. formation and replenishment of CLS/Website repositories/databases/rules/filters in the cloud/IS as a result of the technological process and processes of content management/integration/support:
        i. content/referential/annotated/bibliographic repositories/databases/rules/filters CLS/Website;
        ii. semantic NLP tools (rules, dictionaries, knowledge base, for example, ontology) for IIS content, in particular, a base of linguistic rules (grapheme, morphological, lexical, syntactic and semantic);
        iii. database of regular CLS users and their profiles;

iv. service content database of CLS functioning;

b. support of IT interaction of CLS information and resource elements;
c. analysis of statistics/efficiency of CLS functioning and interaction with a permanent audience in certain time intervals;
d. generation of user interaction forecast scenarios with CLS;
e. replenishment/modification of statistical data analysis rules.

CLS/Website repository/database is implemented according to a three-level scheme:

- storage files for storing relevant content;
- knowledge base for IIS/modification/maintenance of this content;
- accumulation and appropriate processing of service content.

The CLS server subsystem is formed from part of the functional components of the management/support/integration/content modules, in particular [73-81]:

- IIS content based on linguistic analysis of requests;
- formation and filling of the cache of information blocks of relevant popular content frequently requested by users and visitors for quick access;
- interactive access to relevant CLS profiles/options;
- analysis of user requests to accumulate content cache;
- storage and accumulation of information blocks in the cloud;
- extraction of cached content from the cloud at the request of the user or its destruction due to the onset of unpopularity;
- replenishment/modernization by the moderator of rules and IIS knowledge base/content analysis as requests of regular users;
- analysis of user requests to generate relevant reports.

3. Models and methods

The annotated CLS database is the basis of the Website IIS module. Operational and high-quality IIS in the context of current content ensures its high relevance for the CLS user [81-84]. The use of annotated $<L$ in the IIS module helps to implement effective IIS-relevant content without information noise (Fig. 3). CLS should provide [85-91]:

- generation of Webpage according to the template and content of the Website;
- preservation and maintenance of cache/filling of Web page/Website according to the needs of the target audience;
- provision of prompt access to the Website for all types of users.

Mashup-IS S_{MSS} consists in forming a set of integrated content from Internet resources according to the needs of the target audience and specific user requests for convenient navigation on the Website/Web page [81-91]:

$$S_{MSS} =< X, W, Q, Y, \varpi, \rho >, \quad (1)$$

where X is a set of simultaneously integrated content from Internet resources W, Q is user requests to Website/Webpage Mashup-IS S_{MSS}, Y is a set of relevant content as a result of IIS at the request of a user/visitor of Website/Webpage Mashup-IS S_{MSS}; ϖ is an operator in the integration of content from Internet resources W and ρ is a navigation operator in databases/data/content/filters storage.

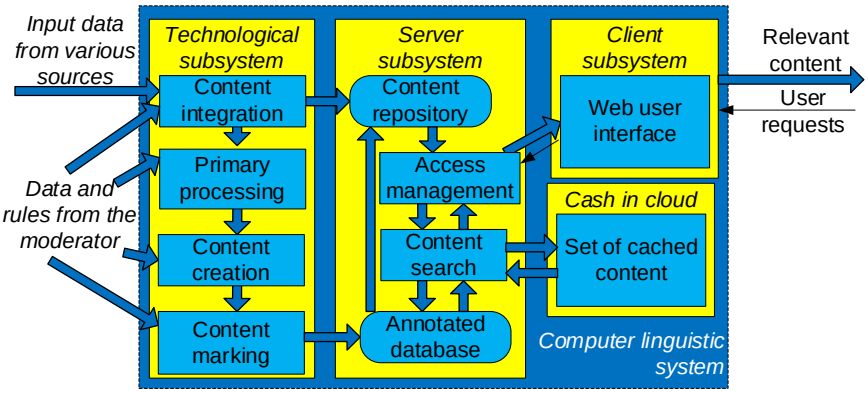

Figure 3: Schematic model of a typical CLS

The integration of a set of data X from various sources W, including the Website, consists of combining them according to the appropriate collection of conditions U_{ϖ} in one Website or web page to use different types of content while preserving its main features, presentation characteristics and the possibility of further processing :

$$X = \varpi(W, U_{\varpi}). \quad (2)$$

The integration should provide the user of the Website/Webpage Mashup-IS S_{MSS} to perceive the integrated content as a single information space using large DS, including the cloud, and high-quality/operational IIS of relevant content upon request according to the collection of IIS conditions U_{ρ}:

$$Y = \rho(Q, X, U_{\rho}). \quad (3)$$

Convenient navigation in the Website/Webpage Mashup-IS S_{MSS} contributes to the realization of the possibility of supporting the user to search for relevant and relevant content for him throughout the available IS information space with the greatest completeness and accuracy with the least expenditure of effort on his part. CLS is a specialized IS, DSS or multi-agent system for solving a specific NLP problem based on a set of integrated content from different sources according to the needs of the target audience and specific user requests for convenient navigation on the Website/Webpage, taking into account the statistics of the CLS operation, history of actions and personal profiles of users and history of requests/transitions from IISS. A typical formal CLS model S_{CLS} will be presented as a tuple:

$$S_{CLS} =< X, W, C, K, Y, D, S_{IAC}, M_{LA}, M_{\nu}, M_{\varpi_1}, M_{\varpi_2}, M_{\rho_1}, M_{\rho_2}, M_{\rho_3}, M_{\rho_4}, M_{\upsilon}, \quad (4)$$

$$\nu, \varpi_1, \varpi_2, \rho_1, \rho_2, \rho_3, \rho_4, \upsilon >,$$

where X is input data to CLS from various sources of information W; Y is source-relevant content from CLS as a result of IIS according to user/visitor requests; M_{LA} is a linguistic content analysis module as a component of the IATCS subsystem S_{IAC}; M_ν is a module for generation/modification of the rules of functioning of all modules from the CLS moderator (for example, rules for updating the cache, integration of content from various sources of information, linguistic IIS, etc.); M_{ϖ_1} is the module for filling the unstructured DB with integrated content X; M_{ϖ_2} is filling module of structured DB based on developed integrated content C; M_{ρ_1} is the module for generating results according to visitors' requests; M_{ρ_2} is the module for generating results according to user requests; M_{ρ_3} is a cache processing module for generating reports on popular requests from CLS users; M_{ρ_4} is cache filling/modification module; M_υ is a module for generating statistical results of the functioning of CLS/modules and user activity D; ν is operator of generation/modification of the rules of operation of all modules from the CLS moderator; ϖ_1 is the operator of filling unstructured DB with integrated content X; ϖ_2 is the operator of filling structured DB based on processed integrated content C; ρ_1 is the operator for generating results according to visitors' requests; ρ_2 is the operator for generating results according to CLS user requests; ρ_3 is cache processing operator for generating Y reports on popular requests from CLS users; ρ_4 is CLS cache filling/modification operator with K data; υ is the operator for generating statistical results of CLS/modules functioning and CLS user activity. Fig. 4 shows the general structure of the proposed typical CLS solution of a specific NLP problem based on functionality and interaction with clouds [81-93].

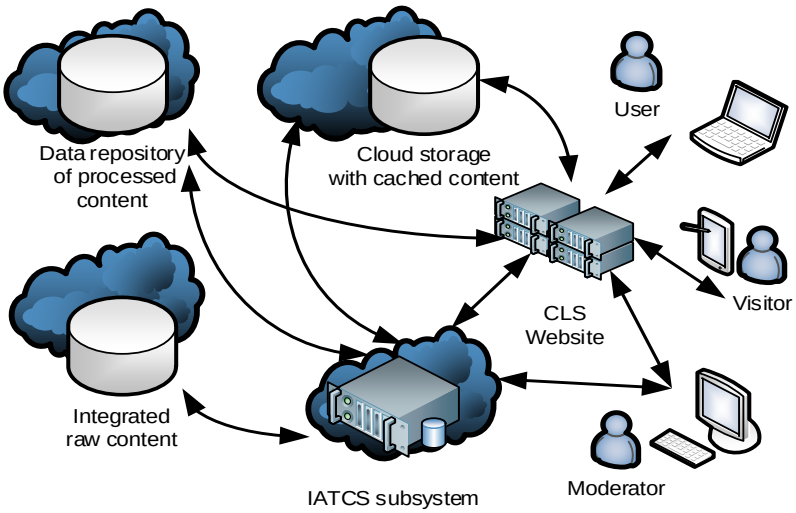

Figure 4: Structure of the process of functioning of a typical CLS

A collection of integrated raw content X is contained in a database based on Non-SQL. A collection of integrated processed C content is contained in a SQL-based repository/database. From C, the filling K for the cloud is formed based on the statistics D

пof popular requests from users for a certain period. Collection D is a specific DB/DS of cached current popular content C to optimize the functioning of CLS based on built-in/modified/additional services in the cloud. These services are the result of the work of the moderator, who updates the caching rules in CLS and/or Website, updates data in the SQL database, IIS/IATCS/ management/ integration/ support of integrated and service content, integration of unprocessed content in Non-SQL database and collection/accumulation of CLS/Website operation statistics.

The basis of the IATCS subsystem is the main NLP processes of CLS.

4. Experiments, results and discussions

4.1. Formal modelling of the main NLP processes of CLS

4.1.1. Formal model of a computer linguistic system for processing Ukrainian-language textual content

Natural languages are determined not by rules, but by the context of the application, which is reconstructed for computer processing. We often identify the meaning of the words used in combination with other interlocutors. The phrase *золота рибка* [zolota rybka] (goldfish) means both a sea creature and a person with a short memory, or a wish-fulfilling creature/person, so the interlocutors must agree to a common understanding of the context. Accordingly, speech/language is limited by mentality, region, society and level of education. Conveying content/meaning is easiest for interlocutors with similar life experiences, education, place of residence, etc. Therefore, automating the understanding of speech to solve a specific NLP problem through the appropriate CLS is a rather complex and painstaking process, especially for synthetic languages, particularly for the Ukrainian language. The general formal model of CLS is given by the collection:

$$S_{LA} =< X, Y, C, D, R, \alpha, \beta, \gamma, \delta, \lambda, o, \iota, \varsigma, \mu >, \quad (5)$$

where X is the input text data array; Y is a tuple of the original processed text according to CLS purpose; C is a set of intermediate content that is processed at the corresponding level in CLS; D is auxiliary dictionaries; R is a set of content processing rules; α is PHA or GA text operator; β is MA text operator; γ is LA content operator; δ is SYA content operator; λ is semantic analysis operator; o is operator of ontological content analysis; ι ia content reference analysis operator; ς is structural content analysis operator; μ is PA content operator.

Compared to formal languages (subject/thing/object), natural languages are more universal, but less formalized. We often use one word to describe several meanings (for example, *краб* [krab] (a crab) is a sea creature, a dish, a nebula in the constellation of Taurus and a cockade on a sailor's cap) depending on the content of the dialogue (for example, a description of the emotions of diving, dinner, a book read, visiting a museum, watching a historical film, etc. only for the word *краб*). To store multiple meanings for each word, the language must be redundant (exceeding the amount of information to transmit/store a message over its entropy). That is, it is not possible to determine in advance the exact

meaning of the content for each association (every linguistic variable is ambiguous by default). Lexical and structural ambiguity is a great achievement of natural language, for example, for generating new ideas, and manifestations of creativity.

Regardless of the NLP task, the process of processing Ukrainian-language texts in arbitrary CLS is presented as a sequence of mandatory operators for meaningful structural analysis of input text content:

information source → *input text* → *grapheme analysis* α (GA) or *phonological analysis* (PHA)→ *morphological analysis* β (MA) → *lexical analysis* γ (LA)→ *syntactic analysis* δ (SYA) → *semantic analysis* λ (SEM) → *structured text content*

Additional operators are analyzed such as pragmatic μ (extraction of knowledge), ontological ο and referential ι (formation of interphrase units). Their application depends on the complexity and purpose of solving the NLP problem.

The main process of linguistic analysis of textual content is presented:

$$Y = \mu \circ o \circ \varsigma \circ \iota \circ \lambda \circ \delta \circ \gamma \circ \beta \circ \alpha, \tag{6}$$

$$Y = \mu(C_\mu, D_\mu, R_\mu, o(C_o, D_o, R_o, \varsigma(C_\varsigma, D_\varsigma, R_\varsigma, \iota(C_\iota, D_\iota, R_\iota, , \lambda(C_\lambda, D_\lambda, R_\lambda, \delta(C_\delta, D_\delta, R_\delta, \gamma(C_\gamma, D_\gamma, R_\gamma, \beta(C_\beta, D_\beta, R_\beta, \alpha(C_\alpha, D_\alpha, R_\alpha, X))))))))), \tag{7}$$

where multiple textual content $C = \{C_\mu, C_o, C_\varsigma, C_\iota, C_\lambda, C_\delta, C_\gamma, C_\beta, C_\alpha\}$, linguistic dictionaries $D = \{D_\mu, D_o, D_\varsigma, D_\iota, D_\lambda, D_\delta, D_\gamma, D_\beta, D_\alpha, \}$ and sets of production/association rules $R = \{R_\mu, R_\iota, R_o, R_\varsigma, R_\lambda, R_\delta, R_\gamma, R_\beta, R_\alpha\}$.

The main linguistic process of processing textual Ukrainian-language information to solve a specific NLP problem consists of nine stages:

Stage 1. Grapheme analysis C_α of textual Ukrainian-language information X:

$$C_\alpha = \alpha(X, D_\alpha, R_\alpha), \tag{8}$$

$$C_\alpha = \alpha_7 \circ \alpha_6 \circ \alpha_5 \circ \alpha_4 \circ \alpha_3 \circ \alpha_2 \circ \alpha_1, \tag{9}$$

where X is the input text data array; α is the GA operator; C_α is grapheme structure of the input text; D_α is grapheme dictionaries and libraries; R_α is grapheme analysis rules; α_1 is OCR operator; α_2 is grapheme parsing operator of the input text X into sections (information blocks), paragraphs and sentences; α_3 is grapheme analysis operator of linguistic chains into separate words; α_4 is the operator for forming a set of unrecognized chains; α_5 is the operator of identification and marking of unrecognized chains as numbers, dates, constant returns, abbreviations, proper and geographical names, etc.; α_6 is the operator for marking non-text strings as special symbols, formulas, figures, tables, etc.; α_7 is an operator for generating a marked linear sequence of words C_α with official signs and connections. GA is replaced by PHA in the case of human speech content recognition.

Stage 2. Morphological analysis β of text content C_β consists in the identification, analysis and determination of the form and structure of words, in particular:

$$C_\beta = \beta(C_\alpha, D_\beta, R_\beta), \tag{10}$$

$$C_\beta = \beta_3 \circ \beta_2 \circ \beta_1, \text{or } C_\beta = \beta_3 \circ \beta_4 \circ \beta_1, \tag{11}$$

where β_1 is the morphological segmentation operator of a graphemically recognized chain of symbols (words/tokens); β_2 is lemmatization operator of lexemes; β_3 is POST operator (marking of parts of speech) for segmented words; β_4 is words stemming operator.

Classical general algorithm of morphological analysis.

Step 1. Morphological segmentation of the input chain of symbols (replacing GA for short English-language messages, and supplementing GA for large corpora of English-language texts, and for Ukrainian-language texts of all kinds - a separate step for marking words in two sets as immediately identifiable (for example, prepositions) or impossible to identify (the noun is not in the nominative case).

Step 2. Lemmatization (reduction to normal form based on dictionary analysis) or stemming – determination of bases (word forms with endings cut off).

Step 3. Identification of the grammatical category of each word and the collection of their corresponding properties in relation to the use in a specific place of the text. (for example, a collection for a noun: gender, case, person, etc.).

Step 4. Formation of a linear sequence of morphological structures.

Stage 3. Lexical analysis γof text content C_γ in the intermediate stage of token sequence analysis for generating a parsing tree at the SYA level:

$$C_\gamma = \gamma(C_\beta, D_\gamma, R_\gamma), \tag{12}$$

$$C'_\gamma = \gamma_2 \circ \gamma_1, , C'_\gamma = \gamma_5 \circ \gamma_4 \circ \gamma_3, \text{ or } C'_\gamma = \gamma_5 \circ \gamma_4. \tag{13}$$

where γ_1 is the Speech Segmentation operator for identification/clarification of words/phrases/tokens after MA or in case of incorrect interpretation during PHA (usually performed in parallel with PHA and MA in a cyclic process); γ_2 is speech recognition operator (SR) or speech-to-text (STT) depending on the content of the NLP task; γ_3 is Optical Character Recognition operator (OCR) as the second part after GA and MA for clarifying incorrect moments of recognition taking into account the recognized neighboring tokens; γ_4 is word tokenization/segmentation operator as data preparation for building a parsing tree in SYA; γ_5 is Text-To-Speech operator (TTS).

Stage 4. The syntactic analysis δ of the text content C_δ consists in building a tree for parsing the dependencies of words in a sequence of tokens based on their categories:

$$C_\delta = \delta(C_\gamma, D_\delta, R_\delta), \text{ or } C_\delta = \delta_3 \circ \delta_2 \circ \delta_1, \tag{14}$$

where δ_1 is the implementation operator of Grammar induction; δ_2 is operator of identification/elimination of boundary ambiguity or sentence violation (Sentence Breaking, Sentence Boundary Disambiguation); δ_3 is operator of syntactic Parsing of phrases/sentences for building a SYA tree.

Stage 5. Semantic analysis λ of textual content C_λ is

$$C_\lambda = \lambda(C_\delta, D_\lambda, R_\lambda), \text{or } C_\lambda = \lambda_2 \circ \lambda_1, \tag{15}$$

where λ_1 is the identification operator of lexical semantics with the generation of a collection of values of each lexeme of the text; λ_2 is the relational semantics identification operator of the interdependencies of the lexeme content of the text.

A classic general algorithm for semantic analysis.

Step 1. Lexemes are compared with meaningful dictionary values.

Step 2. Formation of probabilistic sets for each fragment of text/sentence/phrase with alternative sems, respectively, for lexemes.

Step 3. Preliminary interconnection of the content of tokens into a single structure.

Step 4. Generation of an ordered collection of logical records of superpositions from semantic classes of lexemes and basic lexical functions.

Step 5. Finding/marking inaccuracies, contradictions, incorrectness and ambiguity of the content of the obtained result based on the lexical dictionary.

Semantic analysis is currently not used in most CLS.

Stage 6. Referential analysis ι formation of interphrase units $C_ι$.

$$C_ι = ι(C_λ, D_ι, R_ι). \quad (16)$$

Referential analysis is often a part of semantic analysis. For Slavic languages, when analyzing large text corpora, it is best to take it as a separate stage (for example, to analyze the correspondence of a social group/community in social networks or other dialogues to identify logical meaningful connections between the posts of different participants due to the subjectivity of the speech of each. The classic general algorithm of reference analysis:

Step 1. Contextual analysis of marked fragments of textual content $C_λ$, for example, analysis of the pronoun/conjunction *що* [shcho] (that) analysis depending on the context to separate the center of unity or local references such as *його* [yoho] (his), *який* [yakyy] (which), *цей* [tsey] (this).

Step 2. Actual sentence segmentation of marked fragments of textual content $C_λ$, for identification of thematic structures based on themes/rem.

Step 3. Identification of regular recurrence of context/theme/rheme.

Step 4. Highlighting the duplicated nomination of lexical units of the text.

Step 5. Identification of synonymization of lexical units of the text.

Step 6. Isolation of implications based on situational connections.

Step 7. Identification of the identity of the reference (for example, the comparison of lexical units of the text with the object/subject/phenomenon of dialogue/image).

Stage 7. Structural analysis of ς text content $C_ς$ based on the degree of coincidence of lexical terminological units of the unity of text fragments.

$$C_ς = ς(C_ι, D_ς, R_ς), \text{ or } C_ς = ς(C_λ, D_ς, R_ς). \quad (17)$$

Similarly, referential analysis is often part of SEM for short texts/messages, or not used at all. For large corpora of texts as an additional stage of elimination of marked inaccuracy in SEM. Classical general algorithm of structural analysis.

Step 1. Formation/replenishment of the basic set of rhetorical relations of interphrase units based on the results of reference analysis and/or SEM.

Step 2. Generation of a non-linear network/graph of interphrase units.

Stage 8. Ontological analysis of ο textual content $C_ο$ based on SEM results and reference/structural analyzes if necessary:

$$C_ο = ο(C_ς, D_ο, R_ο), C_ο = ο(C_ι, D_ο, R_ο) \text{ or } C_ο = ο(C_λ, D_ο, R_ο). \quad (18)$$

Stage 9. Pragmatic analysis of μ text content C_μ is used to determine the structure of the text taking into account the context of sentences when forming paragraphs, sections and dialogues. PA is an essential addition to SEM, reference and structural analyzes if they did not contribute to the elimination of marked inaccuracy. In some cases, it is sufficient to apply PA immediately after SEM. It is also an indispensable stage of data preparation for extracting knowledge from text corpora.

$$Y = \mu(C_\mu, D_\mu, R_\mu, C_\lambda, [C_o, C_\varsigma, C_\iota],), \text{or } Y = \mu_2 \circ \mu_1, \quad (19)$$

where μ_1 is the semantics identification operator outside individual sentences/phrases; μ_2 is text processing operator through higher-level NLP applications, for example, to simulate intelligent behavior and apparent understanding of natural language.

4.1.2. Models of grapheme and phonological text analysis in Ukrainian

Depending on the tricky NLP task, grapheme (text analytics) or phonological (speech recognition) analysis of text content is used. PHA consists in the study of the structure, organization, and interpretation of speech sounds X of a specific natural language (Table 1) based on phonemic, phonetic, and prosodic rules of R_α and dictionaries of D_α analogs. GA is a recursive parsing of the text X, taking into account the linguistic features of graphemes of various languages (including non-natural ones, for example, mathematical, programming, pseudo-languages, etc.) based on the rules for recognizing strings of a certain language R_α and dictionaries D_α of reference grapheme models, in particular:

$$C'_\alpha = \alpha(C_\alpha, D_\alpha, R_\alpha, X), C'_\alpha \supseteq C_\alpha. \quad (20)$$

Table 1
Orthographic/phonetic features of the Ukrainian/English language

Unit	Ukrainian language	English language
Sounds	38	44
Letters	32	26
Vowel letters	10	6
Consonant letters	22	20
Vowel sounds	6, there is no division.	12, there are long and short, replacing one sound with another leads to a change in the meaning of the word.
Consonant sounds	32, there are hard and soft, consonants are softened before some vowels (sound *с* [s] in *сіно* [sino] (hay) and *gray* [siryy] (сірий)).	24, there is no division; almost all are pronounced hard before any vowel.
Consonant calls	In such cases, for example *віз* [*вис*] (viz [vis] - cart), *Бог* [*Бох*] (Boh [bokh] - God), *дуб* [*дуп*] (dub [dup]- oak)	They are always pronounced sonically at the end of a word and before voiceless consonants; their stun often leads to a change in the meaning of the word.
Diphthongs (sounds)	There is none	8, some vowels consist of 2 elements pronounced within one syllable.

GA can be an OCR part - encoding/recognizing text on an image into a string of characters for e-submission.

Depending on the tasks, there are the following methods for forming R_α PHA rules:

- General phonology α'_1 (rules of phoneme organization in different languages).
- Descriptive phonology α'_2 (identification of the phoneme of a language or dialect).
- Historical phonology α'_3 (changes in phonemes, language structure during the period).
- Segmental phonology α'_4 (analysis of phonemes, syllables, phonetic words, syntagms and phrases).
- Supersegmental phonology α'_5 (analysis of intonation, tone, stress, rhythm, tempo and pauses).
- Phonetic analysis α'_6 (analysis of the sound structure of the language).
- Phonemic analysis α'_7 of the smallest unit of the phonological level.

GA is an elementary text parsing (Fig. 5) taking into account the features of graphs of different languages and the use of special symbols, objects and marks. A grapheme is an atomic meaningful linguistic (grapheme) unit of a written text (sign, symbol, special symbol, object as a picture, etc.). The purpose of GA is to form a model of the grapheme structure of the input text and generate grapheme rules (regular expressions) for the identification/classification of grapheme units in the sequence of character strings/graphemes and the connections between them. The purpose of the first level of GA – grapheme identification – is to label meaningfully independent sequences of symbols, tokens in these sequences and to determine the main language of the input text content/fragments based on a priori grapheme standards (Fig. 5). The tuple of reference grapheme models is best described by a formal grammar (abbreviations of the criteria are given in Table 2). In parallel with the parsing/identification of graphemes, they are classified/marked according to established rules (Table 3).

Table 2
Criteria for grapheme analysis of input text

No	Abbreviation	Decoding	No	Abbreviation	Decoding
1	*Grammar*	Grammar	35	*Usm*	Ukrainian small letter
2	*Alphabet*	Alphabet	36	*Rcp*	Russian capital letter
3	*Terms*	Term	37	*Rsm*	Russian small letter
4	*Symbol*	Initial character	38	*Cnl*	Consonant letter
5	*PrRules*	Production rules	39	*Vwl*	Vowel letter
6	*Sb*	Symbol	40	*Lcc*	Latin capital consonant letter
7	*Sp*	Space	41	*Lsc*	Latin small consonant letter
8	*Dgt*	Digit	42	*Lcv*	Latin capital vowel letter
9	*Ssb*	Special symbol	43	*Lsv*	Latin small vowel letter
10	*Ssg*	Syntactic sign	44	*Ccc*	Cyrillic capital consonant letter
11	*Ltr*	Letter	45	*Csc*	Cyrillic small consonant letter
12	*Lat*	Latin letter	46	*Ccv*	Cyrillic capital vowel letter
13	*Cyr*	Cyrillic letter	47	*Csv*	Cyrillic small vowel letter
14	*Eng*	English alphabet	48	*Ecc*	English capital consonant letter
15	*Ger*	German alphabet	49	*Esc*	English small consonant letter
16	*Pol*	Polish alphabet	50	*Ecv*	English capital vowel letter
17	*Ukr*	Ukrainian alphabet	51	*Esv*	English small vowel letter
18	*Rus*	Russian alphabet	52	*Gcc*	German capital consonant letter

No	Abbreviation	Decoding	No	Abbreviation	Decoding
19	*Osb*	Official symbol	53	*Gsc*	German small consonant letter
20	*Bsb*	Brackets	54	*Gcv*	German capital vowel letter
21	*Msb*	Mathematical symbol	55	*Gsv*	German small vowel letter
22	*Cpl*	Capital letter	56	*Pcc*	Polish capital consonant letter
23	*Sml*	Small letter	57	*Psc*	Polish small consonant letter
24	*Lcp*	Latin capital letter	58	*Pcv*	Polish capital vowel letter
25	*Lsm*	Latin small letter	59	*Psv*	Polish small vowel letter
26	*Ccp*	Cyrillic capital letter	60	*Ucc*	Ukrainian capital consonant letter
27	*Csm*	Cyrillic small letter	61	*Usc*	Ukrainian small consonant letter
28	*Ecp*	English capital letter	62	*Ucv*	Ukrainian capital vowel letter
29	*Esm*	English small letter	63	*Usv*	Ukrainian small vowel letter
30	*Gcp*	German capital letter	64	*Rcc*	Russian capital consonant letter
31	*Gsm*	German small letter	65	*Rsc*	Russian small consonant letter
32	*Pcp*	Polish letter	66	*Rcv*	Russian capital vowel letter
33	*Psm*	Polish small letter	67	*Rsv*	Russian small vowel letter
34	*Ucp*	Ukrainian capital letter			

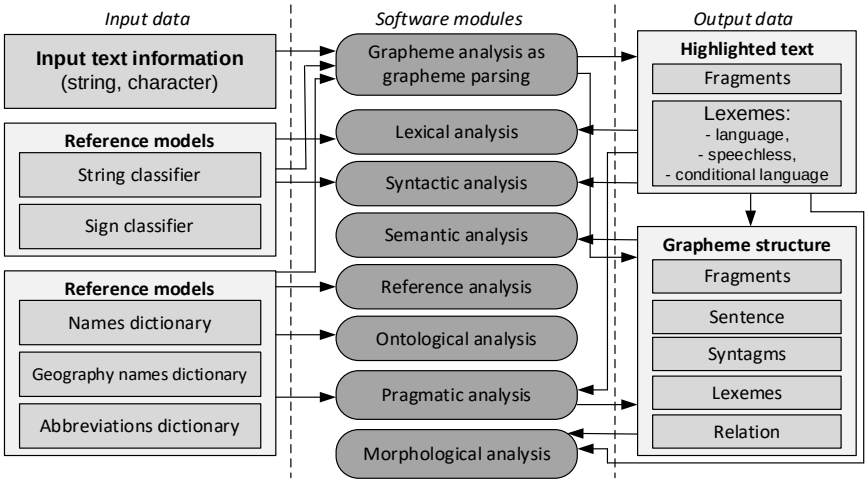

Figure 5: General diagram of the process of grapheme analysis of the input text

Table 3

Rules for identifying graphemes in the form of strings

No	Rule	Transcript	Explanation	Position 1	Position N	All positions
1	EmpStr	empty string	An empty line	–	–	Space
2	FulStr	full string	Full line	Symbol	Symbol	–
3	IncRgt	incomplete right	Incomplete to the right	Symbol	Space	–
4	IncLgt	incomplete left	Incomplete on the left	Space	Symbol	–
5	SmtInc	symmetric incomplete	Symmetrically incomple	Space	Space	–

Let's consider the classical Chomsky $Grammar$ with the $Alphabet$ and $Terms$:

$$Grammar =< Alphabet, Terms, Symbol, PrRules >, \quad (21)$$

$$Alphabet =< Gr, Terms >, \quad (22)$$

$Terms$:=<А, а, Б, б, В, в, Г, г, Ґ, ґ, Д, д, Е, е, Є, є, Ж, ж, З, з, И, и, Й, й, І, і, Ї, ї, К, к, Л, л, М, м, Н, н, О, о, П, п, Р, р, С, с, Т, т, У, у, Ф, ф, Х, х, Ц, ц, Ч, ч, Ш, ш, Щ, щ, ь, Ю, ю, Я, я, Э, э, ы, ъ, ‘, A, a, B, b, C, c, D, d, E, e, F, f, G, g, H, h, I, i, J, j, K, k, L, l, M, m, N, n, O, o, P, p, Q, q, R, r, S, s, T, t, U, u, W, w, V, v, X, x, Y, y, Z, z, Ä, ä, Ö, ö, Ü, ü, Ą, ą, Ć, ć, Ę, ę, Ł, ł, Ń, ń, Ó, ó, Ś, ś, Ź, ź, Ż, ż, ß>.

Table 4 provides a list of grapheme classification/marking rules for reference models (Table 2) according to production rules:

PrRules:=<Symbol→Λ,Symbol→Symbol*Gr*;*Gr*→Λ,*Gr*→*Gr*',*Gr*→*Gr*'*Gr*;*Gr*→*Gr*_*Sp*,*Gr*→*Gr*_*Sb*,*Sb*→*Ssg*,*Sb*→*Ssb*,*Sb*→*Dgt*,*Sb*→*Ltr*;*Sp*→␣,*Ltr*→',*Ltr*→*Vwl*,*Ltr*→*Cnl*,*Ltr*→*Sml*,*Ltr*→*Cpl*,*Ltr*→*Rus*,*Ltr*→*Ukr*;*Ltr*→*Pol*,*Ltr*→*Ger*;*Ltr*→*Eng*,*Ltr*→*Cyr*;*Ltr*→*Lat*,*Ssb*→*Msb*,*Ssb*→*Bsb*,*Ssb*→*Osb*,*Cpl*→*Rcp*,*Cpl*→*Ucp*,*Cpl*→*Pcp*,*Cpl*→*Gcp*,*Cpl*→*Ecp*,*Cpl*→*Ccp*,*Cpl*→*Lcp*,*Sml*→*Rsm*,*Sml*→*Usm*,*Sml*→*Psm*,*Sml*→*Gsm*,*Sml*→*Esm*,*Sml*→*Csm*,*Sml*→*Lsm*,*Lat*→*Lsm*,*Lat*→*Lcp*,*Cyr*→*Csm*,*Cyr*→*Ccp*,*Eng*→*Esm*,*Eng*→*Ecp*,*Ger*→*Gsm*,*Ger*→*Gcp*,*Pol*→*Psm*,*Pol*→*Pcp*,*Ukr*→*Usm*,*Ukr*→*Ucp*,*Rus*→*Rsm*,*Rus*→*Rcp*,*Lcp*→X,*Lcp*→V,*Lcp*→Q,*Lcp*→*Lcv*,*Lcp*→*Lcc*,*Lsm*→x,*Lsm*→v,*Lsm*→q,*Lsm*→*Lsv*,*Lsm*→*Lsc*,*Ccp*→Й,*Ccp*→Ь,*Ccp*→*Csv*,*Ccp*→*Ccc*,*Csm*→й,*Csm*→ь,*Csm*→*Csv*,*Csm*→*Csc*,*Ecp*→X,*Ecp*→V,*Ecp*→Q,*Ecp*→*Lcv*,*Ecp*→*Lcc*,*Esm*→x,*Esm*→v,*Esm*→q,*Esm*→*Lsv*,*Esm*→*Lsc*,*Gcp*→X,*Gcp*→V,*Gcp*→Q,*Gcp*→Ü,*Gcp*→Ö,*Gcp*→Ä,*Gcp*→*Lcv*,*Gcp*→*Lcc*,*Gsm*→x,*Gsm*→v,*Gsm*→q,*Gsm*→ß,*Gsm*→ü,*Gsm*→ö,*Gsm*→ä,*Gsm*→*Lsv*,*Gsm*→*Lsc*,*Pcp*→Ż,*Pcp*→Ź,*Pcp*→Ś,*Pcp*→Ó,*Pcp*→Ń,*Pcp*→Ł,*Pcp*→Ę,*Pcp*→Ć,*Pcp*→Ą,*Pcp*→*Lcv*,*Pcp*→*Lcc*,*Ps*→ż,*Psm*→ź,*Psm*→ś,*Psm*→ó,*Psm*→ń,*Psm*→ł,*Psm*→ę,*Psm*→ć,*Psm*→ą,*Psm*→*Lsv*,*Psm*→*Lsc*,*Ucp*→Ґ,*Ucp*→Ї,*Ucp*→І,*Ucp*→Є,*Ucp*→*Ccv*,*Ucp*→*Ccc*,*Usm*→ґ,*Usm*→ї,*Usm*→і,*Usm*→є,*Usm*→*Csv*,*Usm*→*Csc*,*Rcp*→Ъ,*Rcp*→Э,*Rcp*→Ы,*Rcp*→*Ccv*,*Rcp*→*Ccc*,*Rsm*→ъ,*Rsm*→э,*Rsm*→ы,*Rsm*→*Csv*,*Rsm*→*Csc*,*Lcc*→Z,*Lcc*→W,*Lcc*→T,*Lcc*→S,*Lcc*→R,*Lcc*→P,*Lcc*→N,*Lcc*→M,*Lcc*→L,*Lcc*→K,*Lcc*→J,*Lcc*→H,*Lcc*→G,*Lcc*→F,*Lcc*→D,*Lcc*→C,*Lcc*→B,*Lcv*→Y,*Lcv*→U,*Lcv*→O,*Lcv*→I,*Lcv*→E,*Lcv*→A,*Lsc*→z,*Lsc*→x,*Lsc*→w,*Lsc*→t,*Lsc*→s,*Lsc*→r,*Lsc*→q,*Lsc*→p,*Lsc*→n,*Lsc*→m,*Lsc*→l,*Lsc*→k,*Lsc*→j,*Lsc*→h,*Lsc*→g,*Ls*→f,*Lsc*→d,*Lsc*→c,*Lsc*→b,*Lsv*→y,*Lsv*→v,*Lsv*→u,*Lsv*→o,*Lsv*→i,*Lsv*→e,*Lsv*→a,*Ccc*→Щ,*Ccc*→Ш,*Ccc*→Ч,*Ccc*→Ц,*Ccc*→Х,*Ccc*→Ф,*Ccc*→Т,*Ccc*→С,*Ccc*→Р,*Ccc*→П,*Ccc*→Н,*Ccc*→М,*Ccc*→Л,*Ccc*→К,*Ccc*→З,*Ccc*→Ж,*Ccc*→Д,*Ccc*→Г,*Ccc*→В,*Ccc*→Б,*Csv*→Я,*Csv*→Ю,*Csv*→У,*Csv*→О,*Csv*→И,*Csv*→Е,*Csv*→А,*Csc*→щ,*Csc*→ш,*Csc*→ч,*Csc*→ц,*Csc*→х,*Csc*→ф,*Csc*→т,*Csc*→с,*Csc*→р,*Csc*→п,*Csc*→н,*Csc*→м,*Csc*→л,*Csc*→к,*Csc*→з,*Csc*→ж,*Csc*→д,*Csc*→г,*Csc*→в,*Csc*→б,*Csv*→я,*Csv*→ю,*Csv*→у,*Csv*→о,*Csv*→и,*Csv*→е,*Csv*→а>.

Table 4

Rules for grapheme classification in the form of a sequence of symbols

N	Name with explanation	Grapheme classification rule
1	Input text as a set of characters and spaces;	$Gr := Sb \cup Sp$
2	Space as a terminal character;	$Sp = \{\_\}$
3	A tuple of letters, numbers, special characters and syntax signs;	$Sb := Ltr \cup Dgt \cup Ssb \cup Ssg$
4	A tuple of Latin, Cyrillic and other letters of some European languages and an apostrophe;	$Ltr := Lat \cup Cyr \cup Eng \cup Ger \cup Pol \cup Ukr \cup \cup Rus \cup Cpl \cup Sml \cup Cnl \cup Vwl \setminus \cup \{'\}$
5	A set of numbers;	$Dgt = \{ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 \}$
6	A tuple of all service symbols;	$Ssb := Osb \cup Bsb \cup Msb$
7	A set of syntactic signs;	$Ssg := \{ «, », “, ”, :, -, ?, !\} \cup \{;\} \cup \{,\} \cup \{.\}$
8	The set of capital letters of the corresponding languages;	$Cpl := Lcp \cup Ccp \cup Ecp \cup Gcp \cup Pcp \cup Ucp \cup Rcp$
9	The set of lowercase letters of the corresponding languages;	$Sml := Lsm \cup Csm \cup Esm \cup Gsm \cup Psm \cup Usm \cup Rsm$
10	The set of all Latin letters;	$Lat := Lcp \cup Lsm$
11	Plural of all Cyrillic;	$Cyr := Ccp \cup Csm$
12	The plural of all English letters;	$Eng := Ecp \cup Esm$
13	Set of all German letters;	$Ger := Gcp \cup Gsm$
14	Set of all Polish letters;	$Pol := Pcp \cup Psm$
15	The plural of all Ukrainian letters;	$Ukr := Ucp \cup Usm$
16	Set of all russian letters;	$Rus := Rcp \cup Rsm$
17	Set of service symbols;	$Osb := \{ №, \%, /, @, \#, \$, \&, *, \backslash \}$
18	A set of terminal parentheses;	$Bsb := \{ [,], \{, \}, (,) \}$
19	Set of mathematical symbols;	$Msb := \{ +, <, >, = \}$
20	The set of all consonant letters (uppercase and lowercase) of the corresponding languages;	$Cnl := Ecc \cup Esc \cup Gcc \cup Gsc \cup Pcc \cup Psc \cup \cup Ucc \cup Usc \cup Rcc \cup Rsc$
21	The set of all vowels (uppercase and lowercase) of the respective languages;	$Vwl := Ecv \cup Esv \cup Gcv \cup Gsv \cup Pcv \cup Psv \cup \cup Ucv \cup Usv \cup Rcv \cup Rsv$
22	Multiple Latin capital letters;	$Lcp := Lcc \cup Lcv \cup \{ Q, V, X \}$
23	A set of Latin lowercase letters;	$Lsm := Lsc \cup Lsv \cup \{ q, v, x \}$
24	Multiple Cyrillic capital letters;	$Ccp := Ccc \cup Csv \cup \{ Ь, Й \}$
25	Multiple Cyrillic lowercase letters;	$Csm := Csc \cup Csv \cup \{ ь, й \}$
26	Plural of English capital letters;	$Ecp := Lcc \cup Lcv \cup \{Q, V, X \}$
27	Plural English lowercase letters;	$Esm := Lsc \cup Lsv \cup \{ q, v, x \}$
28	Plural of German capital letters;	$Gcp := Lcc \cup Lcv \cup \{ Ä, Ö, Ü, Q, V, X \}$
29	A set of German lowercase letters;	$Gsm := Lsc \cup Lsv \cup \{ä, ö, ü, ß, q, v, x \}$
30	Plural of Polish capital letters;	$Pcp := Lcc \cup Lcv \cup \{ Ą, Ć, Ę, Ł, Ń, Ó, Ś, Ź, Ż \}$
31	Polish lowercase plural;	$Psm := Lsc \cup Lsv \cup \{ ą, ć, ę, ł, ń, ó, ś, ź, ż \}$
32	Set of Ukrainian capital letters;	$Ucp := Ccc \cup Ccv \cup \{ Є, І, Ї, Ґ \}$
33	Set of Ukrainian lowercase letters;	$Usm := Csc \cup Csv \cup \{ є, і, ї, ґ \}$

N	Name with explanation	Grapheme classification rule
34	The set of russian capital letters;	*Rcp* : = *Ccc* ∪ *Ccv* ∪ { Ы, Э, Ъ }
35	A set of russian lowercase letters;	*Rsm* : = *Csc* ∪ *Csv* ∪ { ы, э, ъ }
36	Latin capital consonant letters;	*Lcc* :={B,C,D,F,G,H,J,K,L,M,N,P,R,S,T,W,Z}
37	Latin capital vowels;	*Lcv* : = { A, E, I, O, U, Y}
38	Terminal Latin small consonant letters;	*Lsc* := {b,c,d,f,g,h,j,k,l,m,n,p,r,s,t,w,x,z}
39	Terminal Latin small vowel letters;	*Lsv*: = { a, e, i, o, u, v, y }
40	The set of terminal Cyrillic capital consonants;	*Ccc* : = { Б, В, Г, Д, Ж, З, К, Л, М, Н, П, Р, С, Т, Ф, Х, Ц, Ч, Ш, Щ }
41	Cyrillic capital vowels;	*Csv* : = { А, Е, И, О, У, Ю, Я }
42	Cyrillic small consonant letters;	*Csc* := {б,в,г,д,ж,з,к,л,м,н,п,р,с,т,ф,х,ц,ч,ш,щ}
43	Terminal Cyrillic lowercase vowels.	*Csv* : = { а, е, и, о, у, ю, я }

$PrRules$ production rules are used to identify, classify and mark meaningful grapheme units of analysis of the input text of content X (words, abbreviations, stable phrases as idioms and metaphors, sentence boundaries and quotations/sarcasms by punctuation, emoticons, geographical and proper names, abbreviations, words with apostrophes, etc.) at the parsing stage, taking into account the language of text fragments. Requirements for identifying a grapheme unit in a sequence of symbols for further morphological analysis of words:

1) the character sequence is easily identified and classified;
2) sequence too large to identify value by dictionary;
3) the sequence is too small to identify many values;
4) the number of grapheme units is too large to split the sample.

4.1.3. Morphological analysis of the Ukrainian language

MA consists in identifying, analyzing and determining the form and structure of words in a natural language text using methods such as Morphological Segmentation β_1, Lemmatization β_2, POST β_3 (marking of parts of speech) and Stemming β_4 (Table 5), in particular:

$$C'_\beta = \beta(C_\beta, D_\beta, R_\beta), \tag{23}$$

where $C'_\beta \subseteq C_\beta$, $C'_\beta = \beta_3 \circ \beta_2 \circ \beta_1$ (enough for English-language short texts of a certain topic) or $C'_\beta = \beta_3 \circ \beta_4 \circ \beta_1$ (for most cases of messages of various topics).

Table 5

Classification of natural language lexeme stemming algorithms

Name	Feature	Disadvantages	Advantages	Example
Lemmatization	1. Identification of parts of speech in a sentence (POST). 2. Application stemming rules according to the part of speech of the word to bring it into normal form. 3. Search in the matching dictionary.	Strong dependence on correct recognition of parts of speech.	High quality and a minimal percentage of errors with correct recognition of part of the language of a word.	The lexemes *вітальне* (adjective) and *пальне* (noun) go through different chains of rules: Rules = { Ending (*льне*) → Cut (*ьне*); Ending (*льне*) → Cut (*е*) }.

Name	Feature	Disadvantages	Advantages	Example
Cutting off suffixes and inflections	Applying the rules for shortening the word to the base (with a prefix) Rules = { Ending (*льне*) → Cut (*ьне*); Ending (*ційним*) → Cut (*ійним*); Ending (*ційний*) → Cut (*ійний*); Ending (*ційне*) → Cut (*ійне*); Ending (*ційна*) → Cut (*ійна*); ……..}.	The presence of false conclusions and distortions of stemming forms (пальне will become пал instead of пальн). Due to the specifics of a specific language, the set of rules is of different levels of complexity and number. There is processing of exceptions, for example, when alternating letters at the base of a word (бігом, біжу). It is necessary to complicate the rules, where a simple cutoff negatively affects the quality of stemming.	Productive and compact, as the number of rules is much smaller than tables with all word forms for all parts of speech, persons, cases, genders, etc.	Word={*національне*} → Stemming={*національ*}; Word={*кульмінаційний*} → Stemming={*кульмінац*}; Word={*приватизаційний*} → Stemming={*приватизац*}; Word={*цивілізаційний*} → Stemming={*цивілізац*}; Word={*інформаційний*} → Stemming={*інформац*}; ……
Separation of prefixes	Along with cutting off the endings and suffixes of the lexeme, separation in the presence of prefixes.	The probability of the formation of words that are opposite in meaning, i.e Word={*незалежний*} → Stemming={*залежн*}.	Significant importance only for some natural languages.	Word={*проголошую, наголошувати, виголошував*} → Stemming={*голошу*}.
IIS according to the table	The dictionary contains all/probable variants of words and their forms after stemming.	Does not work with new words or those whose forms are not presented in the dictionary. Large table sizes for languages with complex morphology (agglutinative, Slavic, including Ukrainian).	Simplicity, speed and convenience of processing exceptions to the rules. For languages with simple morphology (English), the tables are small.	Stemming={*інформац*} → Word={*інформаційний, інформаційна, інформаційне, інформаційним, інформаційними, інформаційних, інформаційні, інформаційній, інформаційнім, інформаційного, безпритульної, інформаційному, інформаційною, інформаційну*}
IIS compliance	Apply knowledge base only with word bases after stemming.	The probability of stemming errors increases with an incorrect description of the rules and formation of the table of endings/inflections	Through the system of rules (matching length of the word and its stem) IIS for the most appropriate form from the KB.	KnowledgeBase={*чорн, чорняв*} → Word={*чорнява*} → Count={4, 6}→ Stemming={*чорнява*}. The algorithm will choose the longer option.
Stemming in different languages	Orientation to competitive language.	The complexity of writing stemming algorithms depends on the features of the language.	The main academic/practical works are devoted only to English.	If English stemming is a simple task, then Ukrainian stemming is several levels more difficult.
Stemming in Ukrainian	Variants of stemming for the Ukrainian language as part of other NLP tasks, but in most cases are commercial projects	There is little research in this direction and there is no free open-source implementation of such algorithms.	Certain steps in this direction have already been taken.	A detailed description of the non-commercial stemming algorithm for Ukrainian is a matter of time.
Stochastic algorithms	They are based on the probability of determining the basis of a word on the basis of KB. Lemmatization has stochastic properties, when a	After processing the word, several variants of the base of the word may appear, from which the algorithm will choose the most likely variant. The probability of stemming errors increases. The most	There is only one logical rule according to which we cut off the last letters from a word. Algorithms have the ability to learn, and the better and larger	Word={*особистість*}→ Stemming={*особист*}→ End={*ість*}; Word={*спогади*}→ Stemming={*спогад*}→ End={*и*}; Word={*дивними*}→ Stemming={*дивн*}→

Name	Feature	Disadvantages	Advantages	Example
	part of speech is determined without taking into account the context in which this word was used in the sentence.	likely part of speech for that word is preferred.	the learning base, the better the result of their work. The knowledge base for these algorithms is a set of logical rules and IIS tables.	End={*ими*}, where End – the result of learning the algorithm, i.e Word(*кияни*) → {End(*ість*) = FALSE, End(*и*) = TRUE, End(*ими*) = FALSE}→ Cut (*и*) or Word(*чуйними*) → {End (*ість*) = FALSE, End (*и*) = TRUE, End (*ими*) = TRUE}→ Cut (*и*) OR Cut (*ими*).
A hybrid approach	A combination of the above algorithms is used.	The probability of stemming errors increases with an incorrect description of the rules and formation of the table of endings	The table does not contain all word forms, but exceptions to the rules that are incorrectly processed by the clipping algorithm.	For example, the algorithm can use the method of cutting off endings and suffixes, but at the first stage, perform IIS on the table.

The best way for Slavic languages:

$$C'_{\beta} = \beta_3 \circ \beta_2 \circ \beta_4 \circ \beta_1. \quad (24)$$

Lematization β_2 is the transformation of a word form into a lemma (normal dictionary form). Usually, during the transformation, a dictionary is used to present words in their actual form. Otherwise, remove only inflections and return to the lemma.

Morphological segmentation β_1 is the division of words into separate morphemes to identify their class (Table 6-8). The complexity is directly proportional to the complexity of the morphology (word structure) of a specific natural language (Table 9-10).

Table 6

Linguistic characteristics of some classes of verb stem morphemes

Verb	Analysis	Basics	Participle
фарбувати(ся)	фарб-ува-ти(-ся)	фарб-$(t, \bar{d}, I, atem, y, ся - \overline{ся})$	фарб-ова-н-ий
усміхнутися	усміх-ну-ти-ся	усміх-$(\bar{t}, d, I, atem, н, ся)$	усміх-н-ен-ий
стогнати	стогн-а-ти	стогн-$(\bar{t}, \bar{d}, I, \tilde{a}, \emptyset, \overline{ся})$	стогн-уч-ий
спитати(ся)	спит-а-ти(-ся)	спит-$(t, \bar{d}, I, a, \emptyset, ся - \overline{ся})$	спит-а-юч-ий
сміятися	смій-а-ти-ся	смій-$(\bar{t}, \bar{d}, I, \tilde{a}, \emptyset, ся)$	смій-уч-ий
розфарбувати(ся)	розфарб-ува-ти(-ся)	розфарб-$(t, d, I, atem, y, ся - \overline{ся})$	розфарб-ова-н-ий
привести(ся)	привес-ти(-ся)	привес-$(t, d, I, atem, \emptyset, ся - \overline{ся})$	приведен-ий
поділити(ся)	поділ-и-ти(-ся)	поділ-$(t, d, \text{II}, \tilde{\imath}, \emptyset, ся - \overline{ся})$	поділ-ен-ий
побудувати(ся)	побуд-ува-ти(-ся)	побуд-$(t, d, I, atem, y, ся - \overline{ся})$	побуд-ова-н-ий
нести(ся)	нес-ти(-ся)	нес-$(t, \bar{d}, I, atem, \emptyset, ся - \overline{ся})$	нес-ен-ий
молоти(ся)	мол-о-ти(-ся)	мол-$(t, d, I, o, \emptyset, ся - \overline{ся})$	мол-о-т-ий і мел-ен-ий
малювати(ся)	мал-ювa-ти(-ся)	мал'-$(t - \bar{t}, d - \bar{d}, I, atem, y, ся - \overline{ся})$	малй-ова-н-ий
любити(ся)	люб-и-ти(-ся)	люб-$(t, \bar{d}, \text{II}, \tilde{\imath}, \emptyset, ся - \overline{ся})$	любл-ен-ий
кохати(ся)	кох-а-ти(-ся)	кох-$(t, \bar{d}, I, a, \emptyset, ся - \overline{ся})$	кох-а-юч-ий
змарніти	змарн-і-ти	змарн-$(\bar{t}, d, I, i, \emptyset, \overline{ся})$	змарн-і-л-ий
запізнюватися → запізнитися	запізн-ювa-ти-ся → запізн-и-ти-ся	запізн-$(\bar{t}, d, I, \tilde{\imath}, \tilde{y}, ся)$	запізн-ювa-н-ий → запізн-ен-ий
досліджувати(ся) → дослідити(ся)	дослідж-ува-ти(-ся) → дослід-и-ти(-ся)	дослідж-$(t, d - \bar{d}, I, \tilde{\imath}, \tilde{y}, ся - \overline{ся})$	дослідж-ува-н-ий → дослідж-ен-ий
втручатися	втруч-а-ти-ся	втруч-$(\bar{t}, \bar{d}, I, a, \emptyset, ся)$	втруч-ен-ий
втратити → втрачати	втрат-и-ти → втрач-а-ти	втрач-$(t, d - \bar{d}, \text{II}, \tilde{\imath}, \emptyset, \overline{ся})$	втрач-ен-ий
вести(ся)	вес-ти(-ся)	вес-$(t - \bar{t}, \bar{d}, I, atem, \emptyset, ся - \overline{ся})$	вед-ен-ий

Verb	Analysis	Basics	Participle
будувати(ся)	буд-ува-ти(-ся)	буд-($t-\bar{t}, d-\bar{d}, I, atem, y, ся-\overline{ся}$)	буд-ова-н-ий
автоматизувати(ся)	автоматиз-ува-ти(-ся)	автоматиз-($t-\bar{t}$, $d-\bar{d}$, I, $atem$, y, $ся-\overline{ся}$)	автоматиз-ова-н-ий

Table 7

Basic rules of formation of Ukrainian participles

Class	Name	Compositional rules
I	General structure rule	1. The word form should include no more than 1 morpheme of each class. 2. Morphemes must be applied in the order of class numbering. 3. Morphemes of classes 1, 4, 5 (stem + suffix + inflection) are mandatory.
II	The rule of incompatibility	A lexeme cannot simultaneously contain: 1. Morphemes of classes 2 and 3 (thematic element and suffix). 2. A stem with the sign $\overline{ся}$ and ending *-ся*. 3. The base with the sign $a/i/o$ and the suffix of the verb with the sign act. 4. A base with the sign $\emptyset$ and a suffix to form verb forms. 5. A stem with the sign d in the absence of a verb suffix and an adjective suffix with the sign $pres$ (present participles are not possible from the perfect form of verbs). 6. The stem with the sign I and without the sign $atem$ and the suffix of the participle with the sign II (verbs of the I conjugation do not allow suffixes of II conjugation). 7. The base with the sign II in the absence of a suffix for the formation of verb forms and the suffix of the verb with the sign I. 8. The suffix for the formation of verb forms and the suffix of the verb with the sign II (the suffix for the formation of verb forms translates any verb in the I conjugation). 9. The stem with the sign $atem$ and the suffix of the participle with the sign II, different from *-ач-/-яч-* (non-TE verbs do not have II declension suffixes, with the exception of *-ач-/-яч-*). 10. A base with the sign $\tilde{\imath}$, $\tilde{a}$ or o, a thematic element and a suffix of a verb beginning with a vowel (if TE is not required for a given base, then it is not used before a suffix of a verb beginning with a vowel). 11. The stem with the sign $atem$ (respectively without the sign atem) and the suffix -юч-/-яч- (respectively -уч-/-ач-), for example, *зітхаю(ть) – зітхаючий*, *співаю(ть) – співаючий*, *квітну(ть) – квітнучий*, *лежа(ть) – лежачий*. These forms are of limited use in the modern Ukrainian language. 12. The verb suffix with the sign act and the inflection with the sign $\bar{f} = o$ (the invariable form of the verb is formed from passive participles by replacing the ending with $\bar{f} = o$, for example, *зроблений – зроблено*, *забитий – забито*, *написаний – написано*, *розглянутий – розглянуто*). 13. A suffix for the formation of perfective/imperfect verb forms of mostly foreign origin and a participle suffix with the signs of act, for example, *наслідувати – наслідуваний*, *гарантувати – гарантований*, *інтенсифікувати – інтенсифікований*, *засохнути – засохлий*, *телеграфувати – телеграфований*, *яровизувати – яровизований*, *організувати – організований*, *організовувати – організований*, *телефонувати – телефонований*, *воєнізувати – воєнізований*, *атакувати – атакований*, *промокнути – промоклий*. 14. Verb suffix and ending -ся (verbs cannot have *-ся*). 15. A stem with the sign $\tilde{\imath}$ (corresponding to the sign $\tilde{a}$) and TE different from *-и(і,ї)-* (corresponding to *-а-/-я-*).
III	The rule of inseparability	The word form must contain: 1. If there is a base with the sign i – TE *-и(і,ї)-*. 2. If there is a base with the sign a – TE *-а-/-я-*. 3. In the presence of a base with the sign of $atem$ and a suffix of a participle beginning with a consonant - either TE, or a suffix for the formation of perfect and imperfect forms of verbs mainly of foreign origin. 4. If there is an infinitive base with the sign *-а* (*-я*), *-ува-* (*-ювa-*), *-овува-* – the suffix *-н-* (*-ий*, *-а*, *-е*, *-і*), is added to it, for example, *посія-(ти) – посіяний*, *чита-(ти) – читаний*, *розпиля-(ти) – розпиляний*, *писа-(ти) – писаний*, *зігна-(ти) – зігнаний*; *загоювати – загоюваний*, *оспівувати – оспівуваний*, *застосовувати – застосовуваний*; the suffix *-ува-* (*-ювa-*), if the stress moves to the first vowel, it changes to *-ова-*, for example, *роздрукува(ти) – роздрукований*, *сформулюва(ти) – сформульований*, *реконструюва(ти) – реконструйований*, *запрограмува(ти) – запрограмований*.

		5. In the presence of a base with the sign *о* – ТЕ *-ор(л)о-* and the possibility of forming parallel forms of participles for infinitive verbs (*колоти – колотий* і *колений*, *пороти – поротий* і *порений, молоти – молотий* і *мелений*). 6. If there is a base with the sign ся – the absence of the particle *ся* in verbs.
IV	Morphological and phonological rules	Morphological rules refer to sequences of graphemes, and necessarily take into account their morphological role. Phonological rules deal simply with sequences of phonemes, regardless of their morphological status. 1. If the base of the infinitive ends in vowels *-и*, *-і* (*-ї*) or consonants, then the suffix *-ен-* (*-єн-*); is the formative; final vowel bases are dropped, and consonants are mostly subject to changes, for example, *втрат-и-ти → втрач-ен-ий*. 2. All verbs in *--отіти* I conjugations that have counterparts in *-отати* (*цокотіти – цокотять*, but *цокотати – цокочуть*): *муркотіти*, *булькотіти*, *тріскотати* etc. Some verbs with a base in *-отати* I conjugations do not have counterparts in *-отіти*, for example, *бельк-ота-ти → бельк-оч-у*, *бельк-оч-уть*, *мурм-ота-ти → мурм-оч-у, мурм-оч-уть*. In order to describe cases such as *бельк-ота-ти* (alternation of *і/а* is impossible) or *цокотіти – цокотати* (alternation of *і/а* is possible, but not necessary) that are not considered here, it is necessary to introduce one more feature of bases: the alternation of *і/а* before *-ти* possible/impossible/obligatory. 3. In the word form containing the suffix *-ува-* (*-ювa-*) and the stress shifts to the first vowel, the suffix changes to *-ова-*, for example, *реконструюва(ти) – реконструйований*, *сформулюва(ти) – сформульований*, *роздрукува(ти) – роздрукований*, *запрограмува(ти) – запрограмований*. 4. In passive verbs *-н-* is not doubled, for example, *намальований*, *зав'язаний*, *зроблений*, *натхнений* etc. 5. In the stems of verbs, the suffix *-ну-* is not preserved when changing form, for example, *стукнути* (d,what to do in future) – *стукати* ($\vec{d}$, what to do now), *крикнути* (d, what to do in future) – *кричати* ($\vec{d}$, what to do now). When forming adjectives, as a rule, it is also dropped, for example, *засохну(ти) – засохлий*, *промокну(ти) – промоклий*. 6. Between two adjacent vowels belonging to different morphemes, *j* appears, for example, *розділ′ + а + йуч + ий*, або *посій + а + н + ий*, *розпил' + а + н + ий*. 7. In case of word change and word formation in the verb forms *г-ж*, *к-ч*, *х-ш*, for example, *берегти – бережу – бережений*, *стерегти – стережу – стережений*. 8. With word change and word formation in the roots of verbs (Table 6). 9. During the formation of verbs, in some cases consonants alternate in personal forms (Table 6). 10. If the base of the infinitive ends in the vowels *-и*, *-і* (*-ї*) or consonants, and the suffix *-ен-* (*-єн-*), is the formative, then the final vowels of the bases are dropped, and the consonants usually undergo changes, for example *пекти – печений*, *заспокоїти – заспокоєний*, *вертіти – верчений*, *пустити – пущений*, *запрягти – запряжений*, *змусити – змушений*, *узгодити – узгоджений*, *вразити – вражений*, *загоїти – загоєний*. Before these suffixes, after labials, *-л-*, appears, for example, *купити – куплений*, *зробити – зроблений*, *вловити – вловлений*, *зломити – зломлений*. 11. Sometimes *-ва-* is lost depending on the tense of the verb, for example, *вбивати – вбити*, *купувати – купити*, але *друкувати – надрукувати, співати – проспівати*. 12. In the word form that has the suffix *-у(ю)ва-*, or the base has the ending *и/а* the root vowel *о* is replaced by *а* in some cases. In order to describe cases not considered here, such as *заспокоїти – заспокоювати* and *заспокоєний – заспокоюваний* (alternation of *о/а* is impossible) or *ломити – ламати* and *ломлений – ламаний* (alternation of *о/а* is possible, but not mandatory), it is necessary to introduce one more feature of the base - alternation *о/а* before *-у(ю)ва-* possible/impossible/obligatory. Alternation in verb roots occurs for such vowels (Table 6). 13. There are rules for using verb suffixes (Table 6). 14. Adverbial suffixes are not doubled, since the stress in adjectives falls on the root (Table 6). 15. Before the suffixes *-е(є)н-*, *-у(ю)ва-*, *-ова-*, *-овува-* the hard final consonants of athematic stems are softened: *д-д′*, *с-с′* etc. 16. Before the suffixes *-е(є)н-*, *-у(ю)ва-*, *-ова-*, *-овува-* the final consonant of the stem *-с′-* is replaced by *-ш-*, and the final consonant *-б′-* by *-бл′-* (similarly, *д′-ж*, *т′-ч*, *в′-вл*, etc.; but our list does not contain stems in *--д′-*, *-т′-*, *-в′-*), for example, *любити – люблю – люблений*, *полюбляти – полюблений*, *висіти – вишу*, *вивішувати – вивішений*; *улюблений*, *робити – роблю*, *роблений*, *виробляю – вироблений*. 17. The unchanging form of the verb is formed from passive participles by replacing the ending with the suffix *-о*, for example, *зроблений – зроблено*, *забитий – забито*, *написаний – написано*, *розглянутий – розглянуто*. It is necessary to use the form with *но*, *-то* instead of passive participles when there is a need to emphasize the action, not the sign, for example, *урок закінчено*, *книжки здано*. 18. The combination *ju* is replaced by *i*.

V	Graphical and orthographic rules	1. The combinations *ja, jy, je, ji* are represented by the letters *я, ю, є, ї,* respectively. 2. The combinations *X'a, X'y, X'e, X'i, X'u* are depicted on the letter as *Хя, Хю, Хє, Хї, Хі* respectively (*X'* is any paired soft consonant).

Table 8

Additional clarifications of morphonological and phonological rules

N	Rule	Example
A	*The basic rules of alternation of consonants in personal forms*	
1	Conjugation (declension) I - consonants change at the end of the stem, if there is an alternation in the 1st person singular – *г-ж, з-ж, к-ч, х-ш, с-ш, т-ч, ст-щ, ск-щ*	*засвистати – засвищу, хотіти – хочу, чесати – чешу, колихати – колишу, мазати – мажу, могти – можу, полоскати – полощу, пе**к**ти – печу – пе**чен**ий*;
2	Conjugation (declension) II - we have sound changes only in the 1st person singular – *д-дж, т-ч, з-ж, с-ш, зд-ждж, ст-щ*	*їздити – їжджу, просити – прошу, возити – вожу, тремтіти – тремчу, водити – воджу, мостити – мощу – мощений*. The only exception is the verb *бігти* (and derivatives: *перебігти, забігти* тощо), in which *г* alternates with *ж* in all personal forms, for example: *бігти – біжу, біжиш, біжить, біжать* (*вибігти – вибіжу, вибіжиш* etc);
B	*Consonant alternation rules in verb roots*	
1	*б-бл*	*полюбляти – полюблений, любити – люблю – люблений, улюблений, робити – роблю, роблений, виробляю – вироблений*;
2	*в-вл*	*ловити – ловлю, виловлювати – виловлений*;
3	*д-дж*	*городити – огороджувати – огороджений*; *городити – загородити – загороджений*;
4	*зд-ждж*	*їздити – їжджу – їжджений*;
5	*з-ж*	*возити – вожу, вивожу – вивезений, лазити – лажу*;
6	*м-мл*	*громити – громлю – погромити – погромлений*;
7	*п-пл*	*терпіти – терплю – терплячий*;
8	*ст-щ*	*розмістити – розміщу, розмістити – розміщувати – розміщений, мастити – мащу, намащую – намащений, мостити – мощу, замощую – замащений*;
9	*с-ш*	*висіти – вишу, вивішувати – вивішений*;
10	*т-д*	*вести – водити, виводити –виведений*;
11	*т-ч*	*летіти – лечу, платити – плачу, сплачувати – сплачений, крутити – кручу – кручений, накручую – накручений*; *платити – сплатити – сплачений*;
12	*ф-фл*	*графити – графлю – графлений, розграфлювати*;
C	*Alternation rules in verb roots for the vowels o and a*	
1	with *a* – repeated, repeated action, imperfect form	*скакати – скакаючий*; *ламати – ламаючий*; *краяти – краючий*; *катати – катаючий*; *хапати – хапаючий*; *ганяти – ганяючий*; *кланятися*; *допомагати*; exception – *вимовляти*; *прощати*; *заспокоювати*; *установлювати*.
2	with *o* – continuous, undivided action or one-time, finished, perfect form	*гонити – гонений*; *схопити – схоплений*; *котити – кочений*; *клонити – клонений*; *кроїти*; *ломити*; *допомогти*; *скочити*; exception – *вимовити*; *простити*; *заспокоїти*; *установити*;
D	*Alternation rules in verb roots for the vowels e (uncased) and i*	
1	with *i* – in prefix verbs of the imperfect form	*використовувати*; *зберігати*; *наріkати*; *випікати*; *замітати*; *вигрібати*; *причіплять and зачіпати*;
2	with *e* – in prefix verbs of the perfect form	*використати*; *зберегти*; *наректи*; *випекти*; *замести*; *вигребти*; *причепити*
3	in verbs with the suffix *-ува-* (*-юва-*) with an accent on the root and in nouns derived from these verbs on *-ння*	*полоскати – виполіскувати – виполіскування, чекати – очікувати – очікування, завертіти – завіртувати – завірчування, брехати – набріхувати – набріхування*, але: *потребувати – потребування, вивершувати – вивершування, прищеплювати – прищеплювання*.
E	*Alternation rules in verb roots for the vowels e (dropped) and i before л, р*	
1	with *и* – in verb roots	*стирати – стертий – стираючий, завмирати – завмираючий, вибирати – вибраний – вибираючий, умирати – умираючий*.
2	with *е* – in verb roots	*завмер – замру – завмираючий, беру – брати – вибраний – вибираючий, вистелю – вислати –висланий – вистеляючий, стер – стертий – стираючий, умерти – умру –умираючий*;
F	*Rules for using adjective suffixes*	
1	*-ян(ий)*	*порівняний*
2	*-ен(ий)*	*завішений, незлічений, нескінчений, неоцінений, куплений*
3	*-ан(ий)*	*сказаний, завішаний, вихований*
G	*Rules for using verb suffixes*	

1	on the first suffix vowel	*-овува-*	*завойовувати – завойовування – завойований*; *перемальовувати – перемальовування –перемальований*.
		-ова-	*підпорядкований,* but *підпорядкувати, підпорядкування*; *мальований,* but *малювати, малювання*; *друкований,* but *друкувати, друкування*; *риштований, риштування,* but *риштувати, риштування*;
2	on the root in derived words and forms (from verbal nouns and verbs)	*-ювa-* *-ува-*	*підбілювати – підбілювання – підбілювaний*; *маршірувати – маршірування*, *бомбувати – бомбування*, *маркувати – маркування*, *вивершувати – вивершування – вивершуваний*, *очікувати – очікування – очікуваний*;

Table 9

Analysis of grammatical/morphological features of Ukrainian/English languages

Linguistic unit	Interpretation of the definition	Language	
		Ukrainian	English
Noun	a name (*Robert*), a person or thing (*a teacher* – вчитель, *a table* – стіл), an action (*a conversation* – розмова).	They have a grammatical gender.	They have no grammatical gender
Noun definition	a noun that is the definition of another - *a stone* bridge (камінний міст).	There is none	It exists
Pronouns	a word used instead of a noun (A boy reads books – *He* reads books). Pronouns are divided into several classes according to their lexical meaning and morphological features	9: personal, inverse, interrogative, relative, possessive, indicative, indicative, indefinite, negative	8: Personal, Possessive, Reflexive, Interrogative, Demonstrative, Relative, Indefinite, Reciprocal
Reflexive pronouns	From possessive pronouns **my, our, your**, etc. by adding endings	*себе*	**The Reflexive Pronouns -** *myself*, *yourself*, *himself*, *herself*, *itself*, *oneself*, *ourselves*, *yourselves*, *themselves*.
Verb	a separate word or phrase describing a state or action (He *loves* his children. Children *play* in the yard Verbs in English, as well as in Ukrainian, mean action (*to go, to build*), state (*to sleep, to rest*), feelings (*to hear, to like*), thinking processes (*to think, to realize*). There are several compound verbs in the English language that have two stems: *to whitewash, to browbeat, to machine-gun*. Many English verbs coincide in form with nouns (rarely with adjectives):	In the Ukrainian language, the verb has 5 typical forms. These forms can be recognized by their characteristic endings: 1) indefinite form (infinitive); 2) personal form: (він) *пиш-***е**, *писа-***в-Ø**, *напиш-***е**, *буде + писа-***ти**, *писати-***ме**, *писа-***в-Ø** + *би*, *хай + пиш-***е**; 3) participle: *пожовті-***л-ий**, *посиві-***л-ий**; *писа-***н-ий**, *підписа-***н-ий**; *залюбл-***ен-ий**, *бач-***ен-ий**; *вими-***т-ий**, *коло-***т-ий**; active participles (writing) are not inherent in the Ukrainian language, this function is performed by descriptive constructions — what (or which) writes.4) impersonal on -но/-то: *написа-***но**, *зробле-*	Modal, emotional, phrasal, irregular. Verbs are simple, derived, complex and compound. Simple verbs consist of one underived stem: *to run, to speak, to go, to try* etc. Derived verbs have suffixes or prefixes: *to organize, rewrite, to discover, to mispronounce*. ***Compound verbs consist of two parts - a verb base and a separated suffix, which are written separately and can be separated by other words***: *to stand up, to sit down, to go away, to put on*та ін. *Sit down, please! Put your cap on!* Compound verbs are very common in the English language. All endings in such verbs are joined to the stem. *He always wakes up at 7 o'clock. I'm writing down your address.*

Linguistic unit	Interpretation of the definition	Language	
		Ukrainian	English
		но,*прожи*-**то**, *випи*-**то**; 5) adverb: *пиш*-**учи**, *любл*-**ячи**, *підписа*-**вши**, *полюби*-**вши**.	
Infinitive	initial form of the verb	without participle: *писа*-**ти**, *говори*-**ти**, *літа*-**ти**, *гримі*-**ти**, *мерзну*-**ти**, *дивува*-**ти**;	as a rule, with the particle to – *to write* (He likes to write letters).
Incorrect forms	those that change according to the usual rules See correct forms.	There is none	Verbs – be *was/wher been* write *wrote written*, degrees of comparison of adjectives/adverbs – good better best.
Participle	it is a form of the verb that means a feature of the subject by action or state and answers the question which? which one which one which? (хмарою повіті, врятована планета, зачарований красою).	Active participles express the sign of an object by its action (палаюче небо). Passive participles express the feature of the subject by action, which is caused by the action of another subject over it (посіяне жито (хтось посіяв)).	An impersonal form of an English verb that has the properties of a verb, an adverb, and an adjective. In Ukrainian, the English participle corresponds to the adverb and the participle.
Present participle-1 time (Present Participle or Participle I).	It has two forms: Present Participle Simple, which corresponds to the Ukrainian present participle. Present Participle Perfect, which corresponds to the Ukrainian present participle and imperfect adverb.	Active participles of the present tense are formed from the base of the present tense of transitive and intransitive verbs of the imperfect form with the help of the suffixes -***уч****(ий),* **-юч***(ий)* for verbs of the 1st conjugation and -***ач****(ий),* **-яч***(ий)* for verbs of the 2nd conjugation -th conjugation (*реве→ревучий, працює→ працюючий*).	-ing form of the verb (reading the book he makes notes – читаючи книгу, він робить позначкиThe Present Participle Simple in the active state is formed by adding the ending –ing to the 1st form of the verb – just like the gerund. In Ukrainian it is translated by an active participle.
Present participle-2 time (Past Participle or Participle II).	t corresponds to the past participle in Ukrainian. speech Passive participles in the Ukrainian language are formed from the base of the infinitive of transitive verbs of the perfect and imperfect forms with the help of the suffixes **-*т****(ий),* **-*н****(ий),* **-ен***(ий),* **-*єн****(ий):* *мити→митий, засіяти→ засіяний, везти→везений, засвоїти→засвоєний.*	Active participles of the past tense are formed from the base of the infinitive only of intransitive verbs of the perfect form with the help of the suffix **-*л****(ий)*:*замерзнути→ замерзлий, побіліти→побілілий.*	the third form of the verb as break-broke-*broken* (a *broken* cup – розбита чашка). The participle of the past tense has only the passive form and is translated as the participle of the past tense into the Ukrainian language.
Adverb	a word that does not change in number or case, indicating how, when, where, where, etc. the action takes place	Він говорить швидко	He speaks *slowly*

Linguistic unit	Interpretation of the definition	Language	
		Ukrainian	English
Quantitative adverbs	an invariable independent part of speech that expresses a sign of action, a state of an object or a sign of quality	answers the question how? where? where? how much? to what extent?	a word that indicates the quantity of something: *many*, *much*, *some*, *any*.
Adjective	describes a person, thing, event, etc. (see degree comparisons of adjectives)	a word that changes in number, gender and case	a word that does not change in number, gender and case (a *tall* boy, a *happy* end, a *long* holiday).
Possessive pronouns	Means belonging to someone or something	Мій, твій, його, її, наш, їхній.	*mine*, *yours*, *his*, *hers*, *ours*, *theirs*
Prepositions	are used before a noun to indicate place, time, direction	They exist, but do not have a literal translation from English	such words as *at*, *in*, *on*, *to*, *under*, *near* (*in* the street, *on* Wednesday, *at* home).
Singular	one thing or person (*a girl*, *a man*, *a child*, *a room*).	There is and is used without an article	There is and is used with the article
Person	the grammatical person of the pronoun	1st person – *я*, *мене*; 2nd person – ти; 3rd person – *він*, *вона*, *воно*, *вони*.	1st person – *I*, *me*; 2nd person – you; 3rd person – *he*, *she*, *it*, *one*, *they*.
Conjunction	v *and* (і, а), *but* (але), *when* (коли), *because* (тому що, бо), connecting sentences.	Йому подобається тяжкий рок, але мені подобається класична музика	He likes hard rock *but* I like classical music.
Time	verb form indicating time	3 forms of the verb depending on the time	12 forms of the verb depending on time (present, past, future).

Table 10

Analysis of syntactic/semantic features of the Ukrainian/English languages

Linguistic unit	Interpretation of the definition	Language	
		Ukrainian	English
Sentence	imperative, exclamatory, interrogative, narrative, negative, affirmative.	Arbitrary order of words	Strict word order
Subject	noun, pronoun or other part of speech that precedes the main verb (predicate).	Машина має двоє дверцят. Ми щасливі.	*A car* has two doors. *We* are happy.
Object	part of speech (noun, pronoun, verb, etc.) that follows the main verb of the sentence (predicate) and answers the question what? whom?	The object is expressed by the same parts of speech as the subject.	The object can be direct, indirect, prepositional. The addition can be expressed by a noun, a pronoun, an infinitive, a gerund, a whole subordinate clause. (I can see *a bus*. They ask me to help.)
Narrative sentence	in which something is affirmed or denied (He speaks English. He doesn't speak English.).	Arbitrary order of words	Strict word order
yes/no question	A general question that needs an answer *yes* (так)/*no* (ні).	Arbitrary order of words	Strict word order
Negative sentence	sentence with a participle not (не) (He *doesn't* speak English. We *don't* like classical music).	Arbitrary order of words	Strict word order
Affirmative sentence	neither negative nor interrogative (He speaks English. We like classical music.).	Arbitrary order of words	Strict word order
Exclamatory sentence	expresses surprise, anger, etc (What a nice day! – Який чудовий день!)	Arbitrary order of words	Strict word order

Linguistic unit	Interpretation of the definition	Language	
		Ukrainian	English
Question	alternative, general yes/no question to the subject, disjunctive, special wh-question.	Arbitrary order of words	Strict word order
Short answer	an answer containing subject+verb-predicate (Who came? – *Mike did*).	Arbitrary order of words	Strict word order
Uncountable nouns	nouns that are not used in the plural: *air* – повітря, *snow* - сніг, *milk* – молоко.	There are exceptions. They have either a singular form or a plural form.	There are exceptions. They have either a singular form or a plural form.
Active state	the action is performed by the subject. See passive condition.	Arbitrary order of words (хлопчик розбив чашку або чашку розбив хлопчик).	Strict word order (A boy *broke* a cup).
Passive state	the action is aimed at the subject	There is none	*be*+ participle – A cup *was broken* (чашку розбито).
Alternative question	a question that offers a choice	Довільний порядок слів (Він говорить англійською чи іспанською? або Говорить він англійською чи іспанською?).	Strict word order (Does he speak English *or* Spanish?).
Question about the subject	a question asking about the subject (formed without an auxiliary verb)	Arbitrary order of words	Strict word order (Who came late? – Jack did.)
Higher degree of comparison	a form of an adjective or adverb used to compare two persons, things, concepts, etc	Two degrees of comparison: higher (смачніший) and highest (найсмачніший)	Three degrees of comparison: ordinary (the Positive Degree), higher (the Comparative Degree) and the highest (the Superlative degree): (*smaller than* – менший ніж, *more expensive* – дорожчий ніж).
A relative subordinate clause	a clause introduced by a relative pronoun	Arbitrary order of words (Це книга, яку я купив вчора або Це книга, яку вчора я купив)	Strict word order (This is the book *which I bought yesterday*).
Relative pronouns	are used to connect the main and subordinate clauses.	what, who, how much, which, whose, which.	*who* (хто, той, що який, котрий); *whom* (кого кому); *that* (який); *which* (котрий, який, хто, що); *whose* (чий, чия, чиє, чиї).
Indicative	The active mode (indicative) means a real action; is the most used. Verbs in the present mood change according to tense.	The meaning of the past and present is real, and the future is hypothetical, so it can have a tinge of the meaning of the invalid mood.	shows that the action is regarded as a real fact in the present, past or future tense.
Plural	more than one (*girls*, *men*, *children*, *rooms*). See singular.	Addition of different inflections depending on the gender of the noun.	Adding the ending s except for exceptions (there is an alternation of letters in the word or the ending en)
Modal verbs	this is a functional-semantic category that expresses the relationship of the content of the statement to reality and the speaker to the content of the statement. Modality is an essential feature of a sentence.	Grammatically, modality is expressed by combining a verb (or other predicate) with modal particles, adverbs, verbs, as well as phrases and sentences. Compare also the different modality in the example: *піди-но, піди,*	*can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, *must*, *ought to*, *need*, *needn't*, *used to*.

Linguistic unit	Interpretation of the definition	Language	
		Ukrainian	English
		ти б пішов, нехай би ти пішов, бодай би ти пішов.	
Sensitive verbs	such verbs as *feel* (почувати), *hear* (чути), *look* (дивитися), *smell* (відчувати запах, нюхати), *sound* (звучати).	There are no syntactic features of use	There are syntactic features of use
The highest degree	form of an adjective or adverb expressing the highest measure	The simple form of the highest degree is formed from the form of the highest degree with the help of the prefix най-. Prefixes like-and what- can be used for reinforcement. The compound form of degrees of comparison of adverbs is formed by adding to a regular adverb: for the higher degree of the words more, less; for the highest degree of the words найбільш, найменш.	*the largest*, *the most important*
Imperative	a verb in the form of an order, instruction, etc	Arbitrary order of words	Strict word order (Open the book! Don't smoke!).
Subjunctive	shows that the speaker considers his action as a real fact, as something permissible, desirable.	Does not exist	Exist
Correct forms	those that change according to the usual rules. See irregular shapes	All verbs have the correct form	Verbs – *play – played – played*, degrees of comparison of adjectives/adverbs – *small – smaller – the smallest*
Possessive	formed with the help of inflections	Or change the word (мама- мамин, тато - татовий)	apostrophe and letter s ('s): 's is added to nouns (common or proper) to indicate belonging to John'*s* father (батько Джона).
Adverbs of degree	such words as *enough* (досить), *fairly* (досить, цілком), *hardly* (ледве, насилу), *quite* (цілком, зовсім, абсолютно, повністю), *rather* (швидше, краще, переважно).	There are	There are (The film was *quite* good.)
Simple time	(past, present, future) expresses short-term action	Does not exist	Exist (The *dance* well. He *lived* in Kyiv).
A divisive question	a short interrogative part that follows the narrative	There is none	Exist. (He speaks English, *doesn't he*? He doesn't speak English, *does he*?).
Complex	(noun, pronoun, etc.) of two or more parts	Хто-небудь, вухогорлоніс.	*schoolboy* (школяр), *somebody* (хтось).
Special wh-question	a question that begins with the interrogative words *who*(*m*), *what*, *when*, *which*, *why*, *where*, *whose*.	Arbitrary order of words	Strict word order
A long time	the form of the verb formed from the auxiliary verb *be* and the semantic verb with the ending *ing be + V-ing*	There is none	indicates that an action is taking place, has taken place or will take place in development. (She *is writing* (вона пише, тобто зараз); She *was writing* (вона писала); She *will be writing* (вона писатиме).
Form -ing	a verb, adjective, or noun ending *-ing*.	There is none	He is *reading* a *boring* book; I like *reading*.
Phrasal verb	a verb used with prepositions	There is none	*Look at* the picture. *Come in*.
Timestamps	adverbs that indicate when an action takes place	Similarly to English	*last year*, *today*, *in* 1994, *on Sunday*

The English language has a significantly simple morphology, especially inflectional morphology, and therefore often completely ignore this task and model all possible word forms accordingly (for example, for *to run* [rʌn] – *run*, *runs*, *ran*, running, for *to work* [wərk] – *work*, *works*, *worked*, *working*) as separate words. In languages such as Turkish or Indian, this approach is not possible because each dictionary entry has thousands of possible word forms. Slavic languages are quite complex and have many endings for one word depending on the case. For example, the verb *бігати* [bihaty] (to run has) a significantly large vocabulary in Slavic languages, so we classify cognate lexemes by parts of speech. Only for the first version of the translation, depending on the structure of the sentence, there may be 36 options for choosing the appropriate verb form in the Ukrainian language and 60 options for the Russian language (Table 11).

Table 11

Forms of the verb *бігати* [bihaty] (to run) for different languages depending on the context

No	English	Ukrainian	Russian	No	English	Ukrainian	Russian
1	run	бігати	бегать	31	running	бігають	бегают
2		біг	бег	32		бігаючи	бегая
3		біжите	бежите	33		біжить	бежит
4		біжать	бегут	34		біжучи	бегучи
5		біжи	беги	35		бігаючи	бегающий
6		біжіть	бегите	36			бегающая
7	I run	бігаю	бегаю	37			бегающее
8		біжу	бегу	38			бегающие
9	let's run	біжимо	бежим	39			бегающего
10		бігаємо, бігаєм	бегаем	40			бегающей
11		біжімо	побежали	41			бегающих
12	you run	бігаєш	бегаешь	42			бегающему
13		бігаєте	бегаете	43			бегающим
14		бігай	бегай	44			бегающую
15		біжиш	бежишь	45			бегающею
16		бігайте	бегайте	46			бегающими
17	runs	бігає	бегает	47			бегающем
18	ran	бігав	бегал	48		бігши	бегавший
19		бігала	бегала	49			бегавшая
20		бігало	бегало	50			бегавшее
21		бігали	бегали	51			бегавшие
22		бігла	бежала	52			бегавшего
23		бігло	бежало	53			бегавшей
24		бігли	бежали	54			бегавших
25	I will run	бігтиму	побегу	55			бегавшему
26	you will run	бігтимеш	побежишь	56			бегавшим
27	you will run	бігтимете	побежите	57			бегавшую
28	will run	бігтиме	побежит	58			бегавшею
29	we will run	бігтимемо	побежим	59			бегавшими
30	will run away	бігтимуть	побегут	60			бегавшем

Then for all of the 13 forms from the Table 13 the word run has 36 variants of analogues in Ukrainian without taking into account the context of a specific sentence - this is more than 450 results. In addition to the common base forms for the verb *бігти* [bihty] (to run), there are more than two dozen, less used words, for example

вбігати [vbihaty] (run in), *вибігати* [vybihaty] (run out), *добігати* [dobihaty] (catch up), *забігати* [zabihaty] (run in), *набігати* [nabihaty] (run over), *оббігати* [obbihaty] (run around), *відбігати* [vidbihaty] (run back), *перебігати* [perebihaty] (run over), *побігати* [pobihaty] (run over), *підбігати* [pidbihaty] (run up), *пробігати* [probihaty] (run over), *збігати* [zbihaty] (run over), *вбігти* [vbihty] (run in), *вибігти* [vybihty] (run out), *добігти* [dobihty] (run over), *забігти*

[zabihty] (run over), *набігти* [nabihty] (run over), *відбігти* [vidbihty] (run back), *перебігти* [perebihty] (run over), *прибігти* [prybihty] (run over), *підбігти* [pidbihty] (run over), *пробігти* [probihty] (run over), *забігати* [zabihaty] (run over), *набігатися* [nabihatysya] (run over), *пробігтися* [probihtysya] (run over), *розбігтися* [rozbihtysya] (run away), *збігтися* [zbihtysya] (converge), *убігати* (run away) [ubihaty], etc.

Each of these words has about 36 variant forms depending on sentence structure and context (over 1000 variant word forms). In addition, the word *run* when translated into Ukrainian can also take on the meaning of a noun, adjective, adverb, adjective or compound word. In addition, the noun and adjective have their own form of declension (7 cases in the Ukrainian language with different inflections and corresponding alternations of letters depending on the rules of the morphology of the language), for example:

- Noun in different genders, for example:

біг [bih] (run), *бігання* [bihannya] (running), *біганина* [bihanyna] (runner), *бігун* [bihun] (runner), *бігунка* [bihunka] (runner), *біженець* [bizhenets'] (refugee), *біженка* [bizhenka] (refugee), *біженство* [bizhenstvo] (refugee), *вибіг* [vybih] (run out), *вибігання* [vybihannya] (running out), *забіг* [zabih] (race), *забігайлівка* [zabihaylivka] (runabout), *забігання* [zabihannya] (short and rare visits), *набіг* [nabih] (raid), *набігання* [nabihannya] (run over), *перебігання* [perebihannya] (running over), *перебіжчик* [perebizhchyk] (defector), *перебіжчиця* [perebizhchytsya] (defector), *побігайчик*[pobihaychyk] (runner), *побігеньки* [pobihen'ky] (runs), *пробіг* [probih] (run), *пробіжка* [probizhka] (scamper), *перебігання* [perebihannya] (running over), *розбіг* [rozbih] (run up), *розбіжка* [rozbizhka] (run off), *збігання* [zbihannya] (coincidence), etc.

- Adjective, for example:

побіжний [pobizhnyy] (cursory), *біговій* [bihoviy] (running), *біженський* [bizhens'kyy] (running), *набіганий* [nabihanyy] (running), *збіганий* [zbihanyy] (running), *вибіганий* [vybihanyy] (running), *забіганий* [zabihanyy] (running), *пробіганий* [probihanyy] (running), *перебіганий*[perebihanyy] (overrunning), *розбіганий* [rozbihanyy] (running), etc.

- Adverbs, for example:

побіжно [pobizhno] (casually), *бігом* [bihom] (by running), *перебіжкою* [perebizhkoyu] (by running), *набігом* [nabihom] (by running), *набігу* [nabihu] (by running), *забігом* [zabihom] (by running), *вибіганням* [vybihannyam] (running out), *забігом* [zabihom] (by running), *набіганням* [nabihannyam] (by running), *перебіганням* [perebihannyam] (by running over), *перебіжчиком* [perebizhchykom] (by a defector), *пробігом* [probihom] (by running), *розбігом* [rozbihom] (by running), *збігом* [zbihom] (by running), etc.

- Adverbs formed from almost every verb, e.g:

вбігати [vbihaty] (run in) – *вбігаючи* [vbihayuchy] (running in), *вибігати* [vybihaty] (run out) – *вибігаючи* [vybihayuchy] (running out), *розбігтися* [rozbihtysya] (run away) – *розбігаючись*[rozbihayuchys'] (running away), etc.

- Complex words of different parts of speech, for example:

Автопробіг [avtoprobih] (car race), *велопробіг* [veloprobih] (bicycle race), *мотопробіг* [motoprobih] (motorcycle race), etc.

And that's just for one word run. This problem is partially solved by the method of marking parts of speech β_3, i.e. parsing into parts of speech or grammatical marking of a word in a corpus (text) taking into account adjacent and related words in a sentence. Many words, especially common ones, can be several parts of speech. For example, *забіг* [zabih] (run) can be a noun in *мій забіг на марафоні [*miy zabih na marafoni] (my marathon run) or a verb in *я забіг у дім* [ya zabih u dim] (I ran home), etc. For most of such cases, identification is possible using Martin Porter's algorithm – stemming β_4 as transforming the word to its base through cutting off inflections, prefixes and suffixes using the appropriate algorithm and rules without a dictionary in contrast to lemmatization β_1 (Table 5):

- IIS according to the table β_4^1;
- Cutting off endings/suffixes according to the rules/trees of endings β_4^2;
- Application of lemmatization rules β_4^3;
- Stochastic algorithms β_4^4;
- A hybrid approach from a combination of the above β_4^5;
- Cutting off prefixes β_4^6;
- IIS correspondence β_4^7.

The choice of a specific algorithm depends on the purpose of the CLS. A combination of these algorithms is usually used for Slavic languages. For example, to process/generate the most difficult word for the Ukrainian language as an adjective and establish it in the required form according to the content of the sentence based on 5 classes of morphemes (Table 12 - Table 13), for example:

розпил'+а+н+ий, *посij+а+н+ий*, *запрограм+ова+н+ий*, *роздрук+ова+н+ий*, *змарн+і+л+ий*, *запізн+і+л+ий*: [rozpyl"+a+n+yy, posij+a+n+yy, zaprohram+ova+n+yy, rozdruk+ova+n+yy, zmarn+i+l+yy, zapizn+i+l+yy];

$$[\text{prefix}] + \{\text{root} + [\text{interfix}]\} + [\text{postsuffix}] + [\text{suffix}] + [\text{ending}]. \quad (25)$$

Table 12

Classes of morphemes for verbs in the Ukrainian language

Class	Name	Example
I	Basis/root	*змарн-*, *роздрук-*, *загоj-*, *заспокоj-*, *розпил'*- etc.; [zmarn-, rozdruk-, zahoj-, zaspokoj-, rozpyl″- etc]
II	Thematic element	*-и(і,ї)-/-а(я)-/-ол(р)о-*; [-y(i,yi)-/-a(ya)-/-ol(r)o-]
III	The postfix of the formation of the perfect and imperfect forms	*-ува-(-ювa-)/-овува-/-ну-*, [-uva-(-yuva-)/-ovuva-/-nu-] for example, *атакувати* [atakuvaty] (attack), *воєнізувати* [voyenizuvaty] (militarize), *гарантувати* [harantuvaty] (guarantee), *інтенсифікувати* [intensyfikuvaty] (intensify), *наслідувати* [nasliduvaty] (emulate), *організовувати* [orhanizovuvaty] (organize), *організувати* [orhanizuvaty] (organize), *телеграфувати* [telehrafuvaty] (telegraph), *телефонувати* [telefonuvaty] (telephone), *яровизувати* [yarovyzuvaty] (vernalize), *засохнути* [zasokhnuty] (dry up), *промокнути* [promoknuty] (soak);
IV	Suffix	*-л-*, *-уч-/-юч-*, *-ач-/-яч-*, *-н-*, *-ен-/-єн-*, *-т-*, *-ова-* etc.; [-l-, -uch-/-yuch-, -ach-/-yach-, -n-, -en-/-yen-, -t-, -ova- etc.]
V	Ending/ inflection	*-а, -і, -е, -у/-ю, -ий, -о* etc. [-a, -i, -e, -u/-yu, -yy, -o etc.]

Table 13

Main linguistic features of sets of morphemes of verb bases

Class	Name	Marking	Example
I	transitivity/ non-transitivity	$t\ /\bar{t}/t-\bar{t}$	*малює* [malyuye] (draws) as type $(t-\bar{t})$
	view of the base/root	$d/\bar{d}/d-\bar{d}$	For $(d-\bar{d})$ the verb is relatively homonymous, for example *автоматизувати* [avtomatyzuvaty] (automate), *досліджувати* [doslidzhuvaty] (investigate), as well as *веліти* [velity] (order), *вінчати* [vinchaty] (crown), *женити* [zhenyty] (marry)
	declension	I/II	Table A.6
II	necessity/ possibility of a thematic element (TE)	$a/i/\tilde{a}/\tilde{\imath}/o$ $/atem$	a – necessary TE *-а-/-я-*(*розпил+я+н+ий*, *чит +а+н+ий*, *пис+а+н+ий*, *леж+а+чий*) [-a-/-ya-(rozpyl+ya+n+yy, chyt +a+n+yy, pys+a+n+yy, lezh+a+chyy)]; i – necessary TE *-и(і,ї)-* (*змарн+і+л+ий*) [-y(i,yi)-(zmarn+i+l+yy)]; $\tilde{a}$ – TE *-а-/-я-* [-a-/-ya-] is possible (*оспів+а+н+ий*, but *оспів+ува+н+ий*) [(ospiv+a+n+yy, but ospiv+uva+n+yy)]; $\tilde{\imath}$ – TE *-и(і,ї)-* [-y(i,yi)-] is possible, but not necessary as *запізн+ювa+н+ий* [zapizn+yuva+n+yy], *запізн+і+л+ий* [zapizn+i+l+yy], *запізн+ ен+ий* [zapizn+ en+yy], *загоїти* [zahoyity] (heal) – *загоювати* [zahoyuvaty] (heal), *змусити* [zmusyty] (force) – *змушувати* [zmushuvaty] (force), *запізнитися* [zapiznytysya] (be late) – *запізнюватися* [zapiznyuvatysya] (to be late), *узгодити* [uzhodyty] (to agree) – *узгоджувати* [uzhodzhuvaty] (to agree), *вирішити* [vyrishyty] (to decide) – *вирішувати* [vyrishuvaty] (to decide); o – is the possibility of forming parallel forms of participles for the bases of infinitive verbs in *-ор(л)о-* [-or(l)o-] (*колоти* [koloty] (pricked) – *колотий* [kolotyy] (pricks)and *колений* [kolenyy] (kneeled), *пороти* [poroty] (torn) – *поротий* [porotyy] (torn) and *порений* [porenyy] (torn), *молоти* [moloty] (grind) – *молотий* [molotyy] (grind) and *мелений* [melenyy] (grind)); $atem$ – impossibility of TE (*вести* [vesty] (lead) – *ведений* [vedenyy] (led));
III-IV	the possibility of attaching a suffix to the base	у/ў/н/∅	у – attachment to the base *-ува-/-ювa-* [-uva-/-yuva-] or *-овува-* [-ovuva-] (*застос+овува+н+ий*, *будувати*) [zastos+ovuva+n+yy, buduvaty]; ў – is possibility to add *-ува-/-ювa-* [uva-/-yuva-] or *-овува-* [-ovuva-] (*заго+ювa+н+ий* or *загої+ти* [zaho+yuva+n+yy or zahoyi+ty], *застосовувати* [zastosovuvaty] (apply), *досліджувати* [doslidzhuvaty] (research), *розпилювати* [rozpylyuvaty] (spray), *зачитувати* [zachytuvaty] (read), *спізнюватися* [spiznyuvatysya] (be late)); н – is the possibility of forming parallel forms of verbs for bases with *-ну-* [-nu-] (*прип+ну+т+ий* [pryp+nu+t+yy], *усуну(ти)* [usunu(ty)] (eliminate) – *усунутий* [usunutyy] (eliminated) and *усунений* [usunenyy] (eliminated); *кину(ти)* [kynu(ty)] (to throw) – *кинутий* [kynutyy] (thrown) and *кинений* [kynenyy] (thrown); *замкну(ти)* [zamknu(ty)] (to close) – *замкнутий* [zamknutyy] (closed) and *замкнений* [zamknenyy] (closed); *верну(ти)* [vernu(ty)] (return) – *вернутий* [vernutyy] (returned) and *вернений* [vernenyy] (returned); *стисну(ти)* [stysnu(ty)] (to compress) – *стиснутий* [stysnutyy] (compressed) and *стиснений* [stysnenyy] (compressed); *зігну(ти)* [zihnu(ty)] (to bend) – *зігнутий* [zihnutyy] (bent) and *зігнений* [zihnenyy] (bent)); ∅ – is the impossibility of a suffix (*запрягти* [zapryahty] (to harness), *пекти* [pekty] (to bake), *опасти* [opasty] (to wear));
V	possibility or need to join −ся [-sya]	ся $/\overline{\text{ся}}$/ся − $\overline{\text{ся}}$	«ся» – is the necessity of adding −ся [-sya] (*розчервонітися* [rozchervonitysya] (to blush), *зажуритися* [zazhurytysya] (become sad), *усміхнутися* [usmikhnutysya] (smile), *намерзатися* [namerzatysya] (freeze), *сміятися* [smiyatysya] (laugh), *втомитися* [vtomytysya] (get tired)), «$\overline{\text{ся}}$» – is the impossibility of adding −ся [-sya] (*стогнати* [stohnaty] (moan)), «ся − $\overline{\text{ся}}$» – is the possibility with and without −ся / $\overline{\text{ся}}$ [-sya] (*купати* [kupaty] (to bathe) – *купатися* [kupatysya] (to bathe)).

For verb suffixes, the type (perfective/imperfect = $d/\bar{d}$), time (present/past = $pres/past$), state (active/passive = act/pas) and declension (I/II/I-II), where (I-II) means that this suffix can be attached to the bases of both I and II of the conjugation (*опалий* [opalyy] (fallen), *зажурений* [zazhurenyy] (sad), *намерзлий* [namerzlyy] (frozen), *розчервонілий* [rozchervonilyy] (reddened), *змарнілий* [zmarnilyy] (wasted), *мерзлий* [merzlyy] (frozen), *промоклий* [promoklyy] (soaked), *засохлий* [zasokhlyy] (dried),

втомлений [vtomlenyy(tired)], *усміхнений* [usmikhnenyy] (smiling)). For verb endings, the full/short form = $f/\bar{f}$ (Table 14).

Table 14

A set of examples of morphemes of all classes with tuples of linguistic features

Class	Example	
I	*втруч*-($\tilde{t}$, $\bar{d}$, *I*, *a*, ∅, *ся*)	*розфарб*-(t, d, *I*, *atem*, у, *ся* − $\overline{ся}$)
	кох-(t, $\bar{d}$, *I*, *a*, ∅, *ся* − $\overline{ся}$)	*вес*-($t-\tilde{t}$, $\bar{d}$, *I*, *atem*, ∅, *ся* − $\overline{ся}$)
	поділ-(t, d, II, $\tilde{\imath}$, ∅, *ся* − $\overline{ся}$)	*буд*-($t-\tilde{t}$, $d-\bar{d}$, *I*, *atem*, у, *ся* − $\overline{ся}$)
	втрач-(t, $d-\bar{d}$, II, $\tilde{\imath}$, ∅, $\overline{ся}$)	*побуд*-(t, d, *I*, *atem*, у, *ся* − $\overline{ся}$)
	смій-($\tilde{t}$, $\bar{d}$, *I*, $\tilde{a}$, ∅, *ся*)	*привес*-(t, d, *I*, *atem*, ∅, *ся* − $\overline{ся}$)
	спит-(t, $\bar{d}$, *I*, *a*, ∅, *ся* − $\overline{ся}$)	*дослідж*-(t, $d-\bar{d}$, *I*, $\tilde{\imath}$, $\tilde{y}$, *ся* − $\overline{ся}$)
	стогн-($\tilde{t}$, $\bar{d}$, *I*, $\tilde{a}$, ∅, $\overline{ся}$)	*автоматиз*-($t-\tilde{t}$, $d-\bar{d}$, *I*, *atem*, у, *ся* − $\overline{ся}$)
	усміх-($\tilde{t}$, d, *I*, *atem*, н, *ся*)	*фарб*-(t, $\bar{d}$, *I*, *atem*, у, *ся* − $\overline{ся}$)
	запізн-($\tilde{t}$, d, *I*, $\tilde{\imath}$, $\tilde{y}$, *ся*)	*люб*-(t, $\bar{d}$, II, $\tilde{\imath}$, ∅, *ся* − $\overline{ся}$)
	мол-(t, d, *I*, *o*, ∅, *ся* − $\overline{ся}$)	*мал'*-($t-\tilde{t}$, $d-\bar{d}$, *I*, *atem*, у, *ся* − $\overline{ся}$)
	змарн-($\tilde{t}$, d, *I*, *i*, ∅, $\overline{ся}$)	*нес*-(t, $\bar{d}$, *I*, *atem*, ∅, *ся* − $\overline{ся}$)
II	*-и(і)- -а(я)- -у(ю)ва- -овува- -ну-*	
III	*-ува-(-ювa-)/-овува-/-ну-*	
IV	*-л-* (I-II,act,$past$,d) *-т-* (I-II,pas,$pres/past$,$d/\bar{d}$)	
	-уч-/-юч- (I,act,$pres$,$\bar{d}$) *-е(є)н-* (I-II,pas,$pres/past$,$d/\bar{d}$)	
	-ач-/-яч- (II,act,$pres$,$\bar{d}$) *-н-* (I-II,pas,$pres/past$,$d/\bar{d}$)	
	-ува-(-ювa-)/-овува-/-ова- (I-II,pas,$pres/past$,$d/\bar{d}$)	
V	f: *-ий -ого -ою -ої -им -ими -ому -их*	
	$\bar{f}$: *-а -е -і -у -о*	

Now we will present the rules for the formation of Ukrainian participles based on the listed morpheme-building forms (Table 6-8). A particular difficulty among these rules is the identification/generation of past participles from imperfect verbs without suffixes. For some cases it is possible (*писаний* [pysanyy] (written), *фарбований* [farbovanyy] (painted),), but for others it is not (*згублений* [hublenyy] (lost)). But there are quite obscure cases (for example, *люблений* [lyublenyy] (loved), *ведений* [vedenyy] (led), *будований* [budovanyy] (built)) when the formation of such verbs is influenced by context and semantics, which cannot be described by the morphological rules of grammar. But in most cases it is possible to generate morphological production rules for the identification/generation of various forms of Ukrainian participles based on Chomsky's formal generative grammar:

$$G = (V, T, S, P), N = V \backslash T, \tag{26}$$

where V is an alphabet, T is a set of terminal elements, $S(S \in V)$ is an initial symbol, P is a set of productions (production rules) of type $X \to Y$, each of which must contain at least 1 non-terminal element from N (Table 15).

Table 15

Non-terminal symbols of the formal generative grammar G_0

No	Symbol	Definition
1	D_K	participle;
2	$D_K(x, y)$	lexeme of a given time and state (x, y are described in rule I);
3	$O'(a_1, a_2, a_3, a_4, a_5, a_6)$	the base of the verb, taking into account TE/suffix if necessary:: a_1 is transitivity $(t/\tilde{t}/t-\tilde{t})$; a_2 is type $(d/\bar{d}/d-\bar{d})$; a_3 is conjugation (I/II); a_4 is thematicity $(a/i/\tilde{a}/\tilde{\imath}/o/atem)$; a_5 is suffix (у/ў/н/∅), a_6 is ending −ся (ся /$\overline{ся}$/*ся* − $\overline{ся}$);

No	Symbol	Definition
4	$O(a_1, a_2, a_3, a_4, a_5)$	base (without ending $-$ся (ся /$\overline{ся}$/$ся - \overline{ся}$));
5	$C(x, y, a_3)$	suffix with signs (x – is state, y – is time, a_3 – is conjugation);
6	$\Phi(и, r, n)$	inflection with signs: $и$ is form (full/abbreviated = $f/\bar{f}$), r is gender category (masculine/feminine/neuter = $m/w/k$), n is number (singular/plural = $s/\overline{s}$).

Production rules for identification/generation of different forms of Ukrainian participles based on Chomsky's formal generative grammar:

I. Substitution rules for the formation of grammatical meanings of the generated verb in textual content in the Ukrainian language.

$$D_K \to D_K(x, y), \tag{27}$$

where $x = (act/pas)$; $y = (pres/past)$, for example,

$$D_K \to D_K(pas, pres), D_K \to D_K(act, pres), \dots \tag{28}$$

II. Substitution rules for generating grammatical values by corresponding morphemes in textual content in the Ukrainian language.

$$\text{II.}\,\mathbf{1}: D_K(act, pres) \to O'\left(t, \bar{d}, a_3\right)C(act, pres, a_3)\Phi, \tag{29}$$

$$\text{II.}\,\mathbf{2}: D_K(act, past) \to O'(\bar{t}, d, a_3)C(act, past, a_3)\Phi, \tag{30}$$

$$\text{II.}\,\mathbf{3}: D_K(pas, pres) \to O'\left(t, d - \bar{d}, a_3\right)C(pas, pres, a_3)\Phi, \tag{31}$$

$$\text{II.}\,\mathbf{4}: D_K(pas, past) \to O'\left(t, d - \bar{d}, a_3\right)C(pas, past, a_3)\Phi, \tag{32}$$

where O, C, Φ are designation of various morphemes without description/identification (Table 15). When forming a description of the corresponding morphemes for reduction, markings of features that acquire different meanings in a specific rule II.i are omitted, for example, $C(act, past)$ is contraction for 2 expressions $C(act, past, a_3)$, so rule II.1 actually consists of many options; $O\left(\bar{d}, a_3\right)$ is an abbreviation for $O\left(a_1, \bar{d}, a_3, a_4, a_5\right)$, where (a_1, a_3, a_4, a_5) take different valid values.

III. Substitution rules for decomposing the verb base (separation of the base of the word and TE/suffix if they are present) in the text in the Ukrainian language.

$$\text{III.}\,\mathbf{1}: O'\left(\overline{atem}\right) \to O\left(\overline{atem}\right)T, \tag{33}$$

where T – is TE; $\overline{atem}$ is is the value of attribute a_4, different from $atem$, i.e$(a/i/\tilde{a}/\tilde{\imath}/o)$.

$$\text{III.}\,\mathbf{2}: O'\left(\overline{d}, \overline{\emptyset}\right)C(x, y) \to O\left(\overline{d}, \overline{\emptyset}\right)C_d C(x, y, I), \tag{34}$$

where C_d is verb suffix; $\overline{\emptyset}$ is any attribute value other than$\emptyset$; x and y must satisfy the following condition: at $x = pas$ it is necessary that$y = pres$.

$$\text{III.}\,\mathbf{3}: O'(atem) \to O(atem), \tag{35}$$

IV. Substitution rules for identification/generation of the TE of the corresponding verb morpheme in the textual content in the Ukrainian language.

$$\text{IV.}\,\mathbf{1}: (\tilde{a})T\alpha \to O(\tilde{a})\zeta, \tag{36}$$

$$\text{IV.}\,\mathbf{2}: O(\tilde{\imath})T\alpha \to O(\tilde{\imath})\zeta, \tag{37}$$

$$\textbf{IV.3}: O(a)T \to O(a)a +, \quad (38)$$

$$\textbf{IV.4}: O(i)T \to O(i)i +, \quad (39)$$

$$\textbf{IV.5}: O(o)T \to O(o)o +, \quad (40)$$

$$\textbf{IV.6}: O(\overline{d}, II, a)TC(act, pres) \to O(\overline{d}, II, a)a + C(act, pres), \quad (41)$$

$$\textbf{IV.7}: O(d - \overline{d}, I, a)TC(pas, pres) \to O(d - \overline{d}, I, a)a + C(pas, pres), \quad (42)$$

$$\textbf{IV.8}: O(d - \overline{d}, I, i)TC(pas, pres) \to O(d - \overline{d}, I, i) + C(pas, pres), \quad (43)$$

$$\textbf{IV.9}: O(\tilde{a}, II)T\beta \to O(\tilde{a}, II)a + \xi, \quad (44)$$

$$\textbf{IV.10}: O(\tilde{\imath}, I)T\beta \to O(\tilde{\imath}, I) + \xi, \quad (45)$$

where ζ and ξ are abbreviated record: ζ is arbitrary vowel, ξ is arbitrary consonant; + is the boundary between morphemes and appears after those that cannot end words.

V. Substitution rules for identification/generation when forming verbs with the corresponding morpheme in textual content in the Ukrainian language.

$$\textbf{V.1}: O(\text{I}, y)C_d \to O(\text{I}, y)\text{ува} +, \quad (46)$$

$$\textbf{V.2}: O(\text{I}, y)C_d \to O(\text{I}, y)\text{ова} +, \quad (47)$$

$$\textbf{V.3}: O(\tilde{y})C_d \to O(\tilde{y}), \quad (48)$$

$$\textbf{V.4}: O(\overline{t}, d, \text{н})C_d \to O(\overline{t}, d, \text{н}) + C(pas, past), \quad (49)$$

$$\textbf{V.5}: O(t, d, \text{н})C_dC(pas, pres) \to O(t, d, \text{н})\text{ну} + C(pas, pres). \quad (50)$$

VI. Substitution rules for the identification/generation of the verb suffix by the corresponding morpheme in the text in the Ukrainian language.

$$\textbf{VI.1}: C(act, past, I - II) \to \text{л} +, \quad (51)$$

$$\textbf{VI.2}: O(atem)C(act, pres, I) \to \text{уч} +, \quad (52)$$

$$\textbf{VI.3}: O(\overline{atem})YC(act, pres, I) \to \text{юч} +, \quad (53)$$

$$\textbf{VI.4}: O(atem)C(act, pres, II) \to \text{ач} +, \quad (54)$$

$$\textbf{VI.5}: O(\overline{atem})YC(act, pres, II) \to \text{яч} +, \quad (55)$$

$$\textbf{VI.6}: C(pas, pres/past, I - II) \to \text{н} +, \quad (56)$$

$$\textbf{VI.7}: C(pas, pres/past, I - II) \to \text{т} +, \quad (57)$$

$$\textbf{VI.8}: O(atem)C(pas, pres/past, I - II) \to \text{ен} +, \quad (58)$$

$$\textbf{VI.9}: O(\overline{atem})YC(pas, pres/past, I - II) \to \text{єн} +, \quad (59)$$

$$\textbf{VI.10}: O(atem)C(pas, pres/past, I - II) \to \text{ува} +, \quad (60)$$

$$\textbf{VI.11}: O(\overline{atem})YC(pas, pres/past, I - II) \to \text{ова} +, \quad (61)$$

$$\textbf{VI.12}: C(pas, pres/past, I - II) \to \text{овува} +, \quad (62)$$

$$\textbf{VI.13}: O(atem)C(pas, pres/past, I - II) \to \text{ова} +, \quad (63)$$

$$\textbf{VI.14}: O(\overline{atem})YC(pas, pres/past, I - II) \to \text{й}\text{ова} +, \quad (64)$$

$$\textbf{VI.15}: O(\overline{atem})X'C(pas, pres/past, I - II) \to X'\text{ьова} +, \quad (65)$$

where Y is any suffix/TE; X' is soft consonant, X is arbitrary consonant.

VII. Substitution rules for choosing the form of the participle ($f/\bar{f}$) and implementing inflection with the corresponding morpheme in the text in the Ukrainian language.

$$\textbf{VII.}\,\textbf{1}\colon \Phi \to \Phi(f), \tag{66}$$
$$\textbf{VII.}\,\textbf{2}\colon \Phi(f,s) \to \text{ого, им, ому}, \tag{67}$$
$$\textbf{VII.}\,\textbf{3}\colon \Phi(f,m) \to \text{ий}, \tag{68}$$
$$\textbf{VII.}\,\textbf{4}\colon \Phi(f,w) \to \text{ою, її}, \tag{69}$$
$$\textbf{VII.}\,\textbf{5}\colon \Phi(f,\bar{s}) \to \text{им, ими, их}, \tag{70}$$
$$\textbf{VII.}\,\textbf{6}\colon \Phi \to \Phi(\overline{f}), \tag{71}$$
$$\textbf{VII.}\,\textbf{7}\colon \Phi(\overline{f},w) \to \text{а}, y, \tag{72}$$
$$\textbf{VII.}\,\textbf{8}\colon \Phi(\overline{f},k) \to \text{е}, \tag{73}$$
$$\textbf{VII.}\,\textbf{9}\colon \Phi(\overline{f},\bar{s}) \to i, \tag{74}$$
$$\textbf{VII.}\,\textbf{10}\colon C(pas)\Phi(\overline{f}) \to \text{о}. \tag{75}$$

VIII. Substitution rules for identification/generation of a base based on a dictionary by the corresponding morpheme in the text in the Ukrainian language.

$$\textbf{VIII.}\,\textbf{1}\colon O(t-\bar{t},\ d-\overline{d},\ I,\ atem,\ \text{у}) \to a\text{втоматиз+},\ \text{буд+},\ \text{мал}'+,\ldots, \tag{76}$$
$$\textbf{VIII.}\,\textbf{2}\colon O(t-\bar{t},\ \overline{d},\ I,\ atem,\ \emptyset) \to \text{вес+},\ \ldots, \tag{77}$$
$$\textbf{VIII.}\,\textbf{3}\colon O(t,\ d-\overline{d},\ II,\ \tilde{\text{і}},\ \emptyset) \to \text{втрач+},\ \ldots, \tag{78}$$
$$\textbf{VIII.}\,\textbf{4}\colon O(\bar{t},\ \overline{d},\ I,\ a,\ \emptyset) \to \text{втруч+},\ \ldots, \tag{79}$$
$$\textbf{VIII.}\,\textbf{5}\colon O(t,\ d-\overline{d},\ I,\ \tilde{\text{і}},\ \tilde{\text{у}}) \to \text{дослідж+},\ \ldots, \tag{80}$$
$$\textbf{VIII.}\,\textbf{6}\colon O(\bar{t},\ d,\ I,\ \tilde{\text{і}},\ \tilde{\text{у}}) \to \text{запізн+},\ \ldots, \tag{81}$$
$$\textbf{VIII.}\,\textbf{7}\colon O(t,\ \overline{d},\ I,\ a,\ \emptyset) \to \text{кох+},\ \ldots, \tag{82}$$
$$\textbf{VIII.}\,\textbf{8}\colon O(t,\ \overline{d},\ II,\ \tilde{\text{і}},\ \emptyset) \to \text{люб+},\ \ldots, \tag{83}$$
$$\textbf{VIII.}\,\textbf{9}\colon O(t,\ \overline{d},\ I,\ atem,\ \emptyset) \to \text{нес+},\ \ldots, \tag{84}$$
$$\textbf{VIII.}\,\textbf{10}\colon O(t,\ d,\ I,\ atem,\ \text{у}) \to \text{побуд+},\ \text{розфарб+},\ \ldots, \tag{85}$$
$$\textbf{VIII.}\,\textbf{11}\colon O(t,\ d,\ II,\ \tilde{\text{і}},\ \emptyset) \to \text{поділ+},\ \ldots, \tag{86}$$
$$\textbf{VIII.}\,\textbf{12}\colon O(t,\ d,\ I,\ atem,\ \emptyset) \to \text{привес+},\ \ldots, \tag{87}$$
$$\textbf{VIII.}\,\textbf{13}\colon O(\tilde{t},\ \overline{d},\ I,\ \tilde{a},\ \emptyset) \to \text{смі}j+,\ \text{стогн+},\ \ldots, \tag{88}$$
$$\textbf{VIII.}\,\textbf{14}\colon O(t,\ \overline{d},\ I,\ a,\ \emptyset) \to \text{спит+},\ \ldots, \tag{89}$$
$$\textbf{VIII.}\,\textbf{15}\colon O(\bar{t},\ d,\ I,\ atem,\ \text{н}) \to \text{усміх+},\ \ldots, \tag{90}$$
$$\textbf{VIII.}\,\textbf{16}\colon O(t,\ \overline{d},\ I,\ atem,\ y) \to \text{фарб+},\ \ldots, \tag{91}$$
$$\textbf{VIII.}\,\textbf{17}\colon O(t,\ d,\ I,\ \text{о},\ \emptyset) \to \text{мол+},\ \ldots, \tag{92}$$
$$\textbf{VIII.}\,\textbf{18}\colon O(\bar{t},\ d,\ I,\ \text{і},\ \emptyset) \to \text{змарн+},\ \ldots, \tag{93}$$

……

IX. Basic morphonological substitution rules for the formation or identification of a participle in textual content in the Ukrainian language.

$$\text{IX. } \mathbf{1}\colon \alpha_1 + \rightarrow \alpha_1 + j\alpha_2, \tag{94}$$

where α_1 and α_2 are arbitrary vowels.

$$\text{IX. } \mathbf{2}\colon j + u \rightarrow i, \tag{95}$$

where *j* is sound designation [*j*] (yot).

$$\text{IX. } \mathbf{3}\colon oZ + C(pas, pres) + Ф \rightarrow aZ + C(act, pres) + Ф, \tag{96}$$

where *Z* is arbitrary sequence no longer than 3 characters (alternation *о*/*а* in the base of the type *скочити/скакати* [skochyty/skakaty] (jump), *ломити/ламати* [lomyty/lamaty] (break), *кроїти/краяти* [kroyity/krayaty] (cut), *клонити/кланятися* [klonyty/klanyatysya] (bow), *котити/катати* [kotyty/kataty] (roll), *схопити/хапати* [skhopyty/khapaty] (grab), *гонити/ганяти* [honyty/hanyaty] (chase), *допомогти/допомагати* [dopomohty/dopomahaty] (help)); group of consonants after that alternating symbol *-о-* (that is, what separates it from ТЕ -а-/-я- before *-у*(*ю*)*ва-*), cannot contain more than 3 letters.

$$\text{IX. } \mathbf{4.1}\colon \text{с}' + W \rightarrow \text{ш} + W, \tag{97}$$

$$\text{IX. } \mathbf{4.2}\colon \text{в}' + W \rightarrow \text{вл}' + W, \tag{98}$$

$$\text{IX. } \mathbf{4.3}\colon \text{б}' + W \rightarrow \text{бл}' + W, \tag{99}$$

$$\text{IX. } \mathbf{4.4}\colon \text{д}' + W \rightarrow \text{дж}' + W, \tag{100}$$

$$\text{IX. } \mathbf{4.5}\colon \text{т}' + W \rightarrow \text{ч} + W, \tag{101}$$

……..

where *W* = *-е*(*є*)*н-*, *-у*(*ю*)*ва-*, *-ова-*, *-овува-*.

$$\text{IX. } \mathbf{5.1}\colon \text{д} + W \rightarrow \text{д}' + W, \tag{102}$$

$$\text{IX. } \mathbf{5.2}\colon \text{с} + W \rightarrow \text{с}' + W, \tag{103}$$

……..

where before the suffix *W*, the hard final consonants of athematic stems are softened: *принес + ти – принес′ + ен+ ий*.

$$\text{IX. } \mathbf{6}\colon \text{нн} + \text{Ф} \rightarrow \text{н} + \text{о}. \tag{104}$$

X. Graphical and orthographic rules of substitution for the formation or identification of a verb in textual content in the Ukrainian language.

$$\text{X. } \mathbf{1.1}\colon j + a \rightarrow я, ja \rightarrow я. \tag{105}$$

$$\text{X. } \mathbf{1.2}\colon j + y \rightarrow ю, jy \rightarrow ю \tag{106}$$

$$\text{X. } \mathbf{1.3}\colon j + e \rightarrow є, je \rightarrow є \tag{107}$$

……..

$$\text{X. } \mathbf{2.1}\colon X' + a \rightarrow X + я, \tag{108}$$

$$\text{X. } \mathbf{2.2}\colon X' + y \rightarrow X + ю, \tag{109}$$

$$\text{X.}\mathbf{2.3}\text{:}\ X^{'} + u \to X + i, \tag{110}$$

$$\text{X.}\mathbf{2.4}\text{:}\ X^{'} + i \to X +, \tag{111}$$

$$\text{X.}\mathbf{2.5}\text{:}\ X^{'} + e \to X + \epsilon, \tag{112}$$

XI. Rules for erasing the indicator of the boundary between verb morphemes in text content in the Ukrainian language.

$$A + B \to AB, \tag{113}$$

where A and B are any morphemes that none of the rules of groups IX-X apply to $A + B$. Such a restriction on A and B prevents untimely destruction of the boundary between morphemes before the application of the corresponding morphological rules. If any morphonological rule can be applied, then it must be applied to prevent the formation of nonsense words, for example, **котаючий* [*kotayuchyy] (*rolling) from *котити* [kotyty] (rolling) or **качений* [*kachenyy] (*rolling) from *катати* [kataty] (rolling). Main properties [404, 882]:

1. Transitivity of derivability. If there is a sequence $A_0, A_1, \ldots, A_n$, in which each i-th chain is *directly derived* from i-1 according to the hypothetical syllogism ($p \to q, q \to r \vdash p \to r$), A_n is derived from A_0, and the sequence $A_0, A_1, \ldots, A_n$ is the *derivation* of $A_n$from A_0 [882].

2. Direct derivability. If there are 2 sequences C and D:

$$C = W_1 F W_2, D = W_1 H W_2, \tag{114}$$

where W_1 and/or W_2 can be empty and the grammar G has the rule $F \to H$, the D is directly derived from C, for example, from the sequence according to rule VI.3$O(t, \bar{d}, I, a, \emptyset,$ ся $- \overline{\text{ся}})a$ + $C(I,act,pres,\bar{d})$ + Ф the sequence is directly derived $O(t, \bar{d}, I, a, \emptyset,$ ся $- \overline{\text{ся}})a$ + *юч* + Ф.

XII. Rules for marking a lexeme as an adjective with multiple features in the corresponding sentence/phrase (gender, tense, case, etc.) in the Ukrainian text.

$$A'_{x,y,z} \to word_{x,y,z}. \tag{115}$$

An example of complete derivation (generation/identification of the verb D_K) in textual content in the Ukrainian language:

(I) $D_K(pas, pres)$

(II.3) $O^{'}(t, d, I, atem, y,$ ся $- \overline{\text{ся}})C(pas, pres, I)\Phi$

(III.3) $O(t, d, I, atem, y)C(pas, pres, I)\Phi$

(VI.13) $O(t, d, I, atem, y)ова + \Phi$

(VII.1) $O(t, d, I, atem, y)ова + \Phi(f)$

(VII.3) $O(t, d, I, atem, y)ова + ий$

(VIII.10) $розфарб + ова + ий$

(XI) *розфарбований*

(XII) $A'_{x,y,z} \to$ *розфарбований*$_{x,y,z}$

An example of dead-end derivation of the verb D_K in the Ukrainian text:

(I) $D_K(act, past)$

(II.2) $O^{'}\left(\overline{t}, d, II, atem, \emptyset,\right.$ ся $\left. - \overline{\text{ся}}\right)C(act, past, II)\Phi$

(Ш.3) $O\left(\overline{t}, d, II, atem, \emptyset\right)C(act, past, II)\Phi$
(VI.1) $O\left(\overline{t}, d, II, atem, \emptyset\right)л + \Phi$
(VII.1) $O\left(\overline{t}, d, II, atem, \emptyset\right)л + \Phi(f)$
(VII.3) $O\left(\overline{t}, d, II, atem, \emptyset\right)л + uй$
(XI) $O\left(\overline{t}, d, II, atem, \emptyset\right)лuй$
(XII) $A'_{x,y,z} \rightarrow лuй_{x,y,z}$
This chain cannot be further generated for the Ukrainian language.

4.1.4. Lexical analysis of the Ukrainian language

LA is preliminary processing of text or speech, transformation of a chain of symbols into a sequence of tokens (tuples of symbols according to appropriate patterns):

$$C'_{\gamma} = \gamma(C_{\gamma}, D_{\gamma}, R_{\gamma}, C_{\beta}), \tag{116}$$

where $C'_{\gamma} \subseteq C_{\gamma}, C'_{\gamma} = \gamma_2 \circ \gamma_1$ (enough to transform the sound series of speech into printed text) or $C'_{\gamma} = \gamma_5 \circ \gamma_4 \circ \gamma_3$ (to transform scanned text into speech), but is sufficient for most cases:

$$C'_{\gamma} = \gamma_5 \circ \gamma_4. \tag{117}$$

1. Speech Segmentation γ_1 is the division of the sound stream of human speech into separate words. In human conversation or speech, pauses between consecutive words are almost unidentifiable, so segmentation is a necessary task for speech recognition. In most spoken languages, sounds as consecutive letters are combined with each other in the process of coarticulation, so converting an analog signal into discrete symbols is a rather complex NLP process in CLS systems for technical implementation.
2. Speech Recognition γ_2 (SR) or Speech-To-Text (STT) is transformation of the speech signal into e-text. Different people pronounce words in each language with different stresses, speeds, and intonations. CLS should recognize a wide range of input unstructured data as identical to a single equivalent.
3. Optical Character Recognition γ_3 (OCR) is translation of scanned handwritten or printed text after GA and MA into a sequence of codes for e-submission with correction of simple errors based on statistics and theory probabilities of using a sequence of N-grams/words/endings instead of obscure symbols.
4. Tokenization or Word Segment γ_4 is the demarcation and categorization of sections of the chain of input symbols for MA. For English or Ukrainian languages, this is a fairly trivial situation, since words are usually separated by spaces (problems only if there are stylistic and grammatical errors in the text). However, some written languages such as Chinese, Korean, Japanese, and Thai do not mark word boundaries in this way. Then tokenization is an important task based on vocabulary and word morphology. Sometimes the method is used to form a Bag-of-words model (BOW) in Data Mining.

5. Text-To-Speech γ_5 (TTS) is transformation of handwritten, typewritten or printed text into a speech signal for oral presentation, for example, for people with visual impairments.

4.1.5. Syntactic analysis of the Ukrainian language

SYA is the basis of semantic analysis:

$$C'_\lambda = \delta(C_\delta, D_\delta, R_\delta), \tag{118}$$

where $C'_\lambda \subseteq C_\lambda$, $C'_\lambda = \delta_3 \circ \delta_2 \circ \delta_1$:

1. Grammar induction δ_1 is generation of formal grammar to describe language syntax.
2. Sentence Breaking or Sentence Boundary Disambiguation δ_2 is analysis of the presence/absence of appropriate punctuation marks and the text between them (a dot not only marks the end of a sentence, but also a contraction).
3. Parsing δ_3 is the generation of sentences from the input sequence of symbols of the parsing tree (grammatical analysis) for the analysis of the grammatical structure according to the given formal grammar (Table 16). There are hundreds or thousands of analyzes for a typical sentence, most of which are absolutely meaningless to a native speaker. There are two main types of parsing: Dependency Parsing δ_3^1 and Constituency Parsing δ_3^2, or in the form of some combination of these methods δ_3^3. Dependency parsing focuses on the relationship between words in a sentence (primary objects and predicates), and constituent parsing focuses on building a SYA tree using probabilistic context-free (stochastic) grammar (PCFG).

For example, when parsing/generating sentences/phrases, the choice of case inflection of a specific word in the Ukrainian language directly depends on the type of base and part of speech, in particular, for noun groups taking into account the context (Table 9-10, Table 16):

$$R \rightarrow E_1 X \rightarrow E_1 C_1 C_2, R \rightarrow E_2 X \rightarrow E_2 C_3 C_4, \tag{119}$$

A set of linguistic units X of one type, together with a set of linguistic units E_1 of another type, are transformed in a different way $C_1 C_2$, than together with a set of linguistic units E_2 of the third type – $C_3 C_4$. Without taking into account the context, a more fractional classification should be introduced:

$$R \rightarrow X_1 \rightarrow C_1 C_2, R \rightarrow X_2 \rightarrow C_3 C_4, \tag{120}$$

Table 16
Rules for formulating Ukrainian phrases

No	Name of the rule	Rule
I.	Choice of structure R	$R \rightarrow \#\tilde{S}_{x,y,\text{н},w}\ \tilde{V}_{y,\text{тепер},w}\#.$
II.	Noun group	1) $\tilde{V}_{x,y,z,3} \rightarrow \tilde{S}_{x,y,z,3}\ \tilde{S}_{x',y',\text{р},w}$; 2) $\tilde{S}_{x,y,z,3} \rightarrow A_{x,y,z}\ \tilde{S}_{x,y,z,3}$;

No	Name of the rule	Rule
III.	Verb group	3) $K_1\tilde{S}_{x,y,z,w}\ K_2 \to K_1 S^{\text{займ}}_{x,y,z,w} K_2, K_1 \neq A_{x,y,z}, K_2 \neq \tilde{S}_{z'}$; 4) $\tilde{S}_{x,y,z,3} \to S_{x,y,z}$. 1) $\tilde{V}_{y,\text{тепер},w} \to V_{y,\text{тепер},w}\ \tilde{S}_{x',y',\text{зн},w'}\ \tilde{S}_{x'',y'',\text{ор},w''}$; 2) $\tilde{V}_{y,\text{тепер},w} \to V_{y,\text{тепер},w}\ \tilde{S}_{x',y',\text{ор},w'}\ \tilde{S}_{x'',y'',\text{зн},w''}$;
IV.	Word replacement	3) $\tilde{V}_{y,\text{тепер},w} \to V_{y,\text{тепер},w}\ \tilde{S}_{x',y',\text{зн},w'}$; 4) $\tilde{V}_{y,\text{тепер},w} \to V_{y,\text{тепер},w}\ \tilde{S}_{x',y',\text{ор},w'}$. 1) $S_{\text{ч},y,z} \to \text{син}_{y,z}, \ldots$; 2) $S_{\text{ж},y,z} \to \text{посм}i\text{шка}_{y,z}, \ldots$; 3) $S_{\text{сер},y,z} \to \text{щастя}_{y,z}, \ldots$ 4) $S^{\text{займ}}_{\text{х,од},z,1} \to \text{я}_z$; 5) $S^{\text{займ}}_{\text{х,од},z,2} \to \text{ти}_z$; 6) $\text{А}_{\text{х},y,z} \to \text{веселий}_{\text{х},y,z}, \text{безмежний}_{\text{х},y,z}, \text{м}i\text{й}_{\text{х},y,z}, \text{тв}i\text{й}_{\text{х},y,z}, \ldots$; 7) $V_{y,\text{тепер},w} \to \text{наповнити}_{y,\text{тепер},w}, \ldots$

Each record is a rules set, for example, II.1 forms 648 rules: $\tilde{S}_{\text{ч,од,н,3}} \to \tilde{S}_{\text{ч,од,н,3}}\ \tilde{S}_{\text{ч,од,р,1}}$; $\tilde{S}_{\text{ч,од,р,3}} \to \tilde{S}_{\text{ч,од,р,3}}\ \tilde{S}_{\text{ч,од,р,1}}$; ...; $\tilde{S}_{\text{сер,мн,м,3}} \to \tilde{S}_{\text{сер,мн,м,3}}\ \tilde{S}_{\text{сер,мн,род,3}}$ (Table 17). To generate a sentence tree in Ukrainian, the inflections agreement rules corresponding are used (Fig. 6-Fig. 7).

Table 17

Designation of grammatical categories of the noun/verb group and constituents

Type	Description
	Noun group
Noun group /$\tilde{N}$	adjective/*A*, noun/*N*, pronoun/$N^{займ}$;
Numeric /*ЧЛ*	singular/*од*, plural/*мн*;
Genus /*РД*	male/*ч*, female/*ж*, medium/*с*;
Case /*ВД*	nominative/*н*, generic/*р*, dative/*д*, accusative/*з*, instrumental/*о*, local/*м*, vocative/*к*;
Person/ОС	1-st/*1*, 2-nd/*2*, 3-th/*3*.
	Verb group
Verb group /$\tilde{R}$	verb/*R*, within the Noun group adjective /*A*, noun/*N*;
Numeric /*ЧЛ*	singular/*од*, plural/*мн*;
Genus /*РД*	male/*ч*, female/*ж*, medium/*с*;
Person/*ОС*	male/*ч*, female/*ж*, medium/*с*;
Time/*ЧС*	present/*тп*, past/*мн*, future/*мб*.

Each step is an expansion of a symbol of the sequence (for example, $\tilde{R}_{од,тп,3}$ – $R_{од,тп,3}\ \check{S}_{ч,од,з,1}\ \check{S}_{с,од,о,3}$) or replacement (so, $\check{S}_{ч,од,з,1}$ – $S^{\text{займ}}_{\text{ч,од,з,1}}$). For such deployment, it is necessary to form more detailed types of linguistic units to take into account the location in the context of the sentence, for example:

$$Word_{\text{мн,род}} \to O^i F_{\text{мн,род}}, Word_{мн,род} \to O^i F^i_{мн,род}, \quad (121)$$

, , (4.1)

where $Word$ is the word form of a sentence, O^i is the base of a word of type i ($i = 1,2,3,\ldots$), $F_{\text{мн,род}}$ is the inflection of the genitive plural in Ukrainian languages, for example:

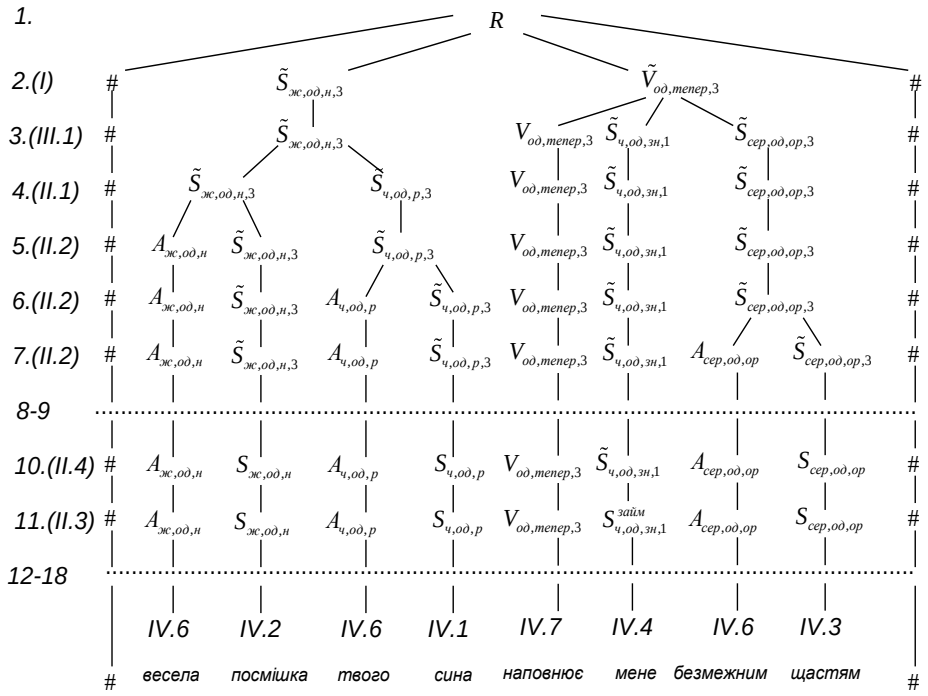

Figure 6: An example of grammar with phrase structure

1. S
2. (I) $\#\tilde{S}_{ж,од,н,3}\ \tilde{R}_{од,тп,3}\#$
3. (III.1) $\#\tilde{S}_{ж,од,н,3}\ R_{од,тп,3}\tilde{S}_{ч,од,з,1}\tilde{S}_{с,од,о,3}\#$
4. (II.1) $\#\tilde{S}_{ж,од,н,3}\ \tilde{S}_{ч,од,р,3}R_{од,тп,3}\tilde{S}_{ч,од,з,1}\tilde{S}_{с,од,о,3}\#$
5. (II.2) $\#A_{ж,од,н}\tilde{S}_{ж,од,н,3}\ \tilde{S}_{ч,од,р,3}R_{од,тп,3}\tilde{S}_{ч,од,з,1}\tilde{S}_{с,од,о,3}\#$
6. (II.2) $\#A_{ж,од,н}\tilde{S}_{ж,од,н,3}\ A_{ч,од,р}\tilde{S}_{ч,од,р,3}R_{од,тп,3}\tilde{S}_{ч,од,з,1}\tilde{S}_{с,од,о,3}\#$
7. (II.2) $\#A_{ж,од,н}\tilde{S}_{ж,од,н,3}\ A_{ч,од,р}\tilde{S}_{ч,од,р,3}R_{од,тп,3}\tilde{S}_{ч,од,з,1}A_{с,од,о}\tilde{S}_{с,од,о,3}\#$
8. (II.4) $\#A_{ж,од,н}S_{ж,од,н}\ A_{ч,од,р}\tilde{S}_{ч,од,р,3}R_{од,тп,3}\tilde{S}_{ч,од,з,1}A_{с,од,о}\tilde{S}_{с,од,о,3}\#$
9. (II.) $\#A_{ж,од,н}S_{ж,од,н}\ A_{ч,од,р}S_{ч,од,р}R_{од,тп,3}\tilde{S}_{ч,од,з,1}A_{с,од,о}\tilde{S}_{с,од,о,3}\#$
10. (II.4) $\#A_{ж,од,н}S_{ж,од,н}\ A_{ч,од,р}S_{ч,од,р}R_{од,тп,3}\tilde{S}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
11. (II.3) $\#A_{ж,од,н}S_{ж,од,н}\ A_{ч,од,р}S_{ч,од,р}R_{од,тп,3}S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
12. (IV.6) $\#$*весела* $S_{ж,од,н}\ A_{ч,од,р}S_{ч,од,р}R_{од,тп,3}S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
13. (IV.2) $\#$*весела посмішка* $A_{ч,од,р}S_{ч,од,р}R_{од,тп,3}S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
14. (IV.6) $\#$*весела посмішка твого* $S_{ч,од,р}R_{од,тп,3}S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
15. (IV.1) $\#$*весела посмішка твого сина* $R_{од,тп,3}S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
16. (IV.7) $\#$*весела посмішка твого сина наповнює* $S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
17. (IV.6) $\#$*весела посмішка твого сина наповнює мене* $A_{с,од,о}S_{с,од,о}\#$
18. (IV.6) $\#$*весела посмішка твого сина наповнює мене безмежним* $S_{с,од,о}\#$
19. (IV.3) $\#$*весела посмішка твого сина наповнює мене безмежним щастям*#

Figure 7: An example of deriving a sentence of a given structural scheme

$$O^1F_{\text{мн,род}} \to O^i\text{ів}(\text{друз} - \text{ів}), F^1_{мн,род} \to ів, \tag{122}$$

$$O^2F_{\text{мн,род}} \to O^i\text{ок}(\text{іграш} - \text{ок}), F^2_{мн,род} \to ок, \tag{123}$$

$$O^3F_{\text{мн,род}} \to O^i\text{ей}(\text{діт} - \text{ей}), F^3_{мн,род} \to ей, \tag{124}$$

$$O^4F_{\text{мн,род}} \to O^i\text{их}(\text{знайом} - \text{их}), F^4_{мн,род} \to их, \quad (125)$$

$$O^5F_{\text{мн,род}} \to O^i(\text{машин} -), F^5_{мн,род} \to \Lambda. \quad (126)$$

……

The choice of the case of the direct complement in a word form or sentence depends in the Ukrainian language on the presence/absence of negation:

$$\tilde{V} \to V^i\tilde{S}_d, \tilde{V} \to V^i\tilde{S}^1_d, \quad (127)$$

$$\tilde{V} \to \neg V^i\tilde{S}_d, \tilde{V} \to \neg V^i\tilde{S}^2_d. \quad (128)$$

where $\tilde{V}$ is a verb group in a sentence, V^i is a transitive verb in a verb group, $\tilde{S}_d$ is a direct object, $\tilde{S}$ is a noun group, ¬ is a negation, in particular, for sentences, *школяр пише есе* [shkolyar pyshe ese] (the student writes an essay) and *школяр не пише есе* [shkolyar ne pyshe ese] (the student does not write an essay) with relevant conclusions:

$$XV\tilde{S}_d \to \underset{X\neq\neg x}{XV^i\tilde{S}_з}, \tilde{S}^1_d \to \tilde{S}_з \text{ or respectively } X\neg V\tilde{S}_d \to \underset{X\neq\neg x}{X\neg V^i\tilde{S}_p}, \tilde{S}^2_d \to \tilde{S}_p.$$

The use of the instrumental subject $\tilde{S}^{sb}$ with a verbal noun depends on the presence of the object $\tilde{S}^{ob}$ (*аналіз змісту системою* [analiz zmistu systemoyu] - content analysis by the system):

$$\tilde{S} \to \tilde{S}'\tilde{S}^{ob}\tilde{S}^{sb}, \tilde{S} \to \tilde{S}'\tilde{S}^{ob}\tilde{S}^{sb^1}, \quad (129)$$

$$\tilde{S} \to \tilde{S}'\tilde{S}^{sb}, \tilde{S} \to \tilde{S}'\tilde{S}^{sb^2}, \quad (130)$$

$$\tilde{S}^{ob}\tilde{S}^{sb} \to \tilde{S}^{ob}\tilde{S}_o, \tilde{S}^{sb^1} \to \tilde{S}'_o. \quad (131)$$

It is impossible to completely abandon the semantics of the context during the correct identification and correction of grammatical and stylistic errors, it is necessary to take into account not only one symbol in the left part of the rules (30)-(40), which ensures the permutation of symbols.

Semantic features are investigated using a set of lexical and linguistic resources in the form of D dictionaries and libraries, T_D dictionary management tools, semantic role marking λ^3_2 and the process of word embeddings $\lambda_з$. Word embedding consists in mapping words, phrases or phrases from the dictionary D into vectors of real numbers E_D for ease of processing, for example, based on Word2Vec.

$$S_f = \lambda^3_2(\lambda_з(D, T_D), E_D). \quad (132)$$

Taking semantics into account significantly simplifies the grammatical tree (Fig. 8). If we connect the symbols (*ancestors*) directly to the final results (*descendants*) during expansion, replacement or rewriting, we get a tree of components, or a syntactic structure. The specified rules (Table 16) are capable of generating other phrases that are not necessarily meaningful, as rules II.1 and II.2 are cyclical. Along with the sequence as *весела посмішка* [vesela posmishka] (funny smile), you can get *весела весела посмішка* [vesela vesela posmishka] (funny funny smile), etc. The number of phrases in natural language must be finite. With the correct construction of the sentence parsing tree (Fig. 6-Fig. 8) and further shortening (Fig. 9), it is possible to match cases.

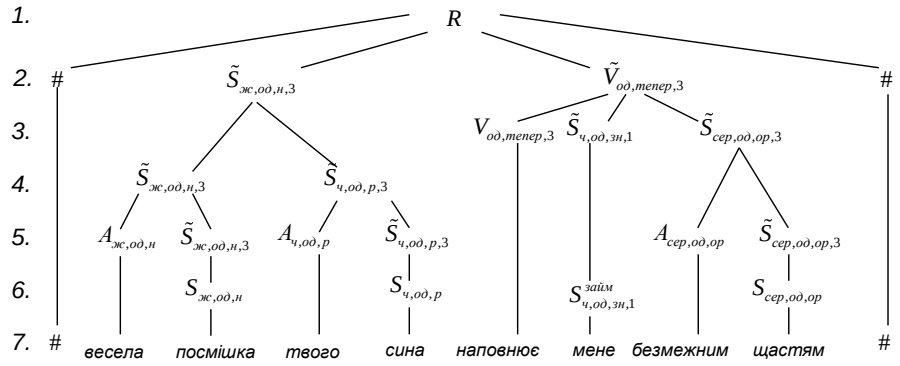

Figure 8: An example of a component tree for context-sensitive grammar

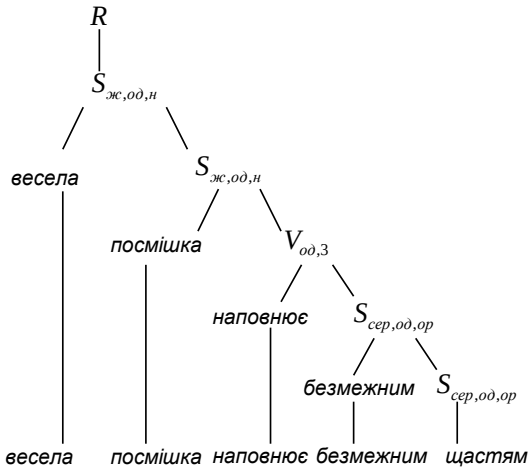

Figure 9: An example of regular grammar (type 3)

The agreement of cases between the linguistic units of the sentence affects the subsequent semantic analysis of the text. For example, in the Ukrainian language it is possible to generate sequences of linguistic units of the type $x_1x_2x_3qx_1'x_2'x_3'$, $x_2x_2x_3qx_2'x_1'x_2'x_3'$, $x_1x_3x_2x_1qx_1'x_3'x_2'x_1'$ etc. (or as XqX'): $\overset{a}{\text{Саша}}$, $\overset{b}{\text{Софія}}$, $\overset{c}{\text{Катя}}$, $\overset{d}{\text{Данило}}$, … – $\overset{a'}{\text{спортсмен}}$, $\overset{b'}{\text{співачка}}$, $\overset{c'}{\text{художниця}}$,$\overset{d}{\text{поет}}$, … *відповідно*. In particular, $x \rightarrow (abcd\ldots)$ is a sequence of proper names; $x' \rightarrow (a'b'c'd'\ldots)$ is a sequence of professions agreed with proper names in the family; q is a punctuation mark.

$$\left.\begin{array}{l} 1.\ R \rightarrow RY_i x_i^{'}, \\ 2.\ x_i^{'}Y_j \rightarrow Y_j x_i^{'}, \\ 3.\ RY_i \rightarrow x_i R, \\ 4.\ R \rightarrow q. \end{array}\right\} i,j = 1,2,3, \tag{133}$$

wherex_i, x_i', q are basic linguistic units; R, Y_i are auxiliary linguistic units; R is the initial symbol as an indicator of the chain generation type. For a more effective study of errors, meta-data analysis of linguistic features and characteristics of the original text is used, in

particular, content genre, presence/absence of dialect, slang, terminology, and the probability of writing by a native speaker or the result of a translation.

4.1.6. Semantic analysis of the Ukrainian language

Semantic analysis is currently not used in most CLS, but with the gradual introduction of AI into the everyday life of the average person, this task should be solved and simplified (Fig. 10). The more complex the grammar of the language, the more difficult it is to conduct CEM:

$$C'_{\mu} = \lambda(C_{\lambda}, D_{\lambda}, R_{\lambda}), \tag{134}$$

where $C'_{\mu} \subseteq C_{\mu}$, $C'_{\mu} = \lambda_2 \circ \lambda_1$.

Linguistic semantics λ_1 (individual words in context) consists of:

$$C''_{\mu} = \lambda_1^6 \circ \lambda_1^5 \circ \lambda_1^4 \circ \lambda_1^3 \circ \lambda_1^2 \circ \lambda_1^1, C''_{\mu} \subseteq C'_{\mu}. \tag{135}$$

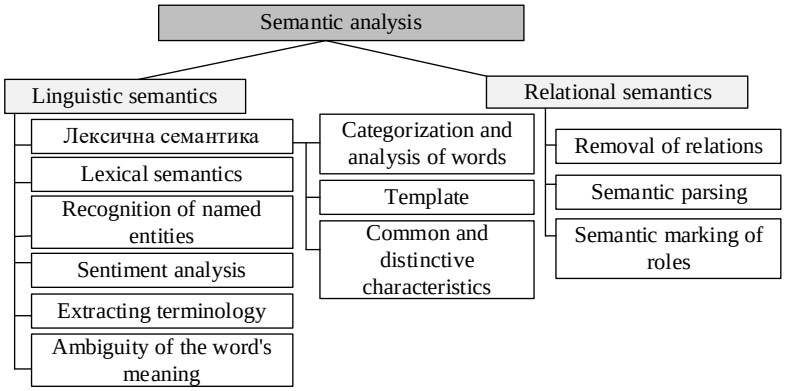

Figure 10: Classification of the main methods of semantic text analysis

1. Lexical Semantics λ_1^1 is analysis of the meanings of individual lexical elements of words/lexemes/morphemes, in contrast to the semantics of sentences, for the construction of a semantic network on based on:
    a. Analysis and categorization of words;
    b. Generation of sets of distinctive/common characteristics in lexico-semantic schemes of different languages;
    c. Formation of a template – the relation of meaning of lexical/phraseological units (content and vocabulary) based on syntax to the content of a specific sentence, analysis of paronyms, homonyms of official and significant words, hyponymy, hyperonymy, antonymy, synonymy.
2. Distributional Semantics λ_1^2 is calculation of the degree of semantic approximation between linguistic units based on their distribution (distribution) in large arrays of linguistic data (text corpora).
3. Word-Sense Disambiguation λ_1^3 (WSD) is definition of a specific meaning of a word from a set of possible ones in a specific sentence to improve the results of CLS work, for example, in discourse analysis, increasing the relevance of IISS, determining

coherence (integrity) of the text (Coherence), Anaphora Resolution and final conclusion (Eng. Inference).

4. Named-Entity Recognition λ_1^4 (NER) or identification of an object entity, fragmentation of an object entity, extraction of an object entity is extraction from unstructured text of information about the presence of certain named entities of the corresponding categories such as dates, percentages, monetary values, quantities, time, geographical locations, names of organizations or people, etc. Having a capital letter at the beginning of a word does not solve this problem - the beginning of a sentence or a line of poetry also begins with a capital letter. In German, all nouns begin with a capital letter. In addition, named entities often include several words, only some of which are capitalized. French, Ukrainian, and Spanish do not capitalize adjective names. In German, all nouns are capitalized. Chinese, Korean, Japanese, or Arabic have no capital letters at all.
5. Terminology Mining λ_1^5 (Terminology Extraction, Term Recognition, Glossary Extraction, Term Extraction) is the automatic extraction of relevant terms from the relevant corpus. One of the first steps towards knowledge domain modeling is the collection of a vocabulary of domain terms as linguistic features of text content. Approaches to term extraction use linguistic processors (marking parts of speech, fragmentation of parts of texts) to extract terminological candidates, that is, syntactically plausible terminological noun phrases or keywords, for example for rubrication.
6. Tonality analysis of the text or multimodal sentiment analysis λ_1^6 (Opinion Mining, Sentiment Analysis) is the detection of a subjective emotionally-colored set of content (positive, negative or neutral) in a text array data of a specific author in relation to the corresponding object, subject, event or phenomenon of thematic subject area. It is useful for identifying trends of public opinion for online marketing, in social networks, forming political propaganda, etc.

Relational semantics λ_2 (semantics of individual sentences) consists of:

$$C'''_{\mu} = \lambda_2^3 \circ \lambda_2^2 \circ \lambda_2^1, C'''_{\mu} \subseteq C'_{\mu}. \tag{136}$$

1. Relationship Extraction λ_2^1 is identification of relationships of nominal entities (for example, family ties, colleagues, enemies, etc.).
2. Semantic Parsing λ_2^2 is presentation of the semantics of a part of the text (usually a sentence) in the form of a logical formalism (DRT parsing, Discourse Representation Theory) or a graph (AMR parsing, Abstract Meaning Representation), for example:

$$\exists a, b, e: f(a, \text{хотіти}_{01}) \wedge f(e, \text{гуляти}_{01}) \wedge f(b, \text{дитина}) \wedge \\ \wedge g_0(a, b) \wedge g_1(a, e) \wedge g_0(e, b). \tag{137}$$

AMR format: (a / хотіти-01 : arg0 (b / дитина) : arg1 (e / гуляти-01 : arg0 b)).

3. Semantic Role Labelling λ_2^3 (Implicit Semantic Role Labeling Below) is assignment of semantic role labels to words or word combinations in a sentence, for example, the role of goal, agent, or result according to the algorithm:

$$C''''_{\mu} = \lambda_2^{35} \circ \lambda_2^{34} \circ \lambda_2^{33} \circ \lambda_2^{32} \circ \lambda_2^{31}, C''''_{\mu} \subseteq C'''_{\mu}. \quad (138)$$

   a. Selection of a sentence or a fragment of text λ_2^{31}.
   b. Definition of semantic predicates in a sentenceλ_2^{32} (verb and noun frames).
   c. Disjunction of defined semantic predicates λ_2^{33}.
   d. Identification of the elements of the defined frame λ_2^{34}.
   e. Classification of the identified elements of the frame λ_2^{35}– assigning a semantic role in the sentence.

4.1.7. Pragmatic analysis of the Ukrainian language

PA is used to determine the structure of the text taking into account the context of sentences when forming paragraphs, sections and dialogues. The main task is the identification of the context of such linguistic units as sentences and the formation of the semantic relationship of these linguistic elements. PA is a mandatory component in AIS for the interpretation of human dialogue, speech and its corresponding analysis (Fig. 11).

$$Y = \mu\left(C_{\mu}, D_{\mu}, R_{\mu}\right), Y = \mu_2 \circ \mu_1. \quad (139)$$

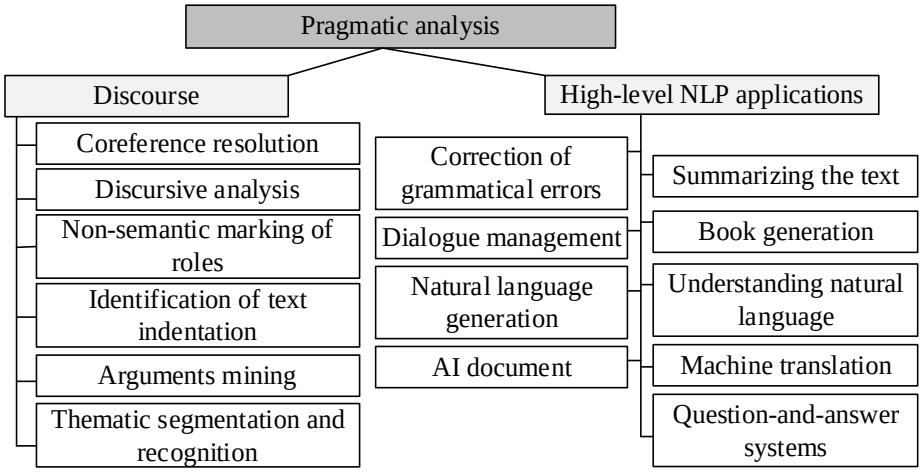

Figure 11: Classification of the main methods of pragmatic text analysis

Discourse μ_1 (semantics outside individual sentences) consists of:

$$Y' = \mu_1^6 \circ \mu_1^5 \circ \mu_1^4 \circ \mu_1^3 \circ \mu_1^2 \circ \mu_1^1, Y' \subseteq Y. \quad (140)$$

1. Coreference Resolution μ_1^1 is identification in the next text of the following words-references or expressions about objects, subjects, phenomena and events, which are

identified in the previous array of text. The resolution of anophora is a relevant example of this task (comparison of pronouns in the subsequent text with nouns or names from the previous fragment of the text). Another task is the identification of bridging relations - expressions that are related to certain objects, subjects, phenomena or events, for example, in the text "she gave warm greetings from her granddaughter Sophia to her grandmother" the expression "warm greetings" is an expressive link and a bridge relation (belong to a specific subject and are assigned to a specific other subject of relations in the corresponding text).

2. Discourse Analysis μ_1^2 consists of the following stages:
    a. Discourse Parsing μ_1^{21} is identification of the discourse structure of a connected text, that is, the characteristics of discourse relationships between sentences (for example, interpretation, improvement, contrast).
    b. Recognizing and Classifying the Speech Acts μ_1^{22} in part of a text fragment (for example, yes-no questions, content questions, assertions, statements, etc.).
3. Implicit Semantic Role Labeling μ_1^3 consists of the following stages:
    a. Semantic marking of SR roles of a text fragment.
    b. Identification of a set of semantic roles of S_N, which are clearly not implemented in the current sentence of the corresponding text.
    c. Classification into appropriate arguments $S_N = S_1 \cup S_2$, which
        i. are clearly implemented in other fragments of text S_1,
        ii. not specified and not implemented throughout text S_2.
    d. Semantic marking of the roles of the first set S1 in the local analyzed fragment of the text.
4. Recognizing Textual Entailment μ_1^4 is comparison of two fragments of text *A* and *B* to form an appropriate conclusion regarding the correspondence of the reality of *A* and:
    a. a logical true derivation of the content of fragment *B* from it,
    b. logical denial of the content *B* of the text fragment from it,
    c. possibilities of text fragmen *B* to be true or false.
5. Topic Segmentation and Recognition μ_1^5 is division of a corresponding text fragment into segments of different topics and identification of the topics of these segments.
6. Argument Mining μ_1^6 is automatic extraction and identification of argumentative structures from the text, which include the premise, conclusions, argumentation scheme and the relationship between the main and auxiliary argument or the main and counterargument in the discourse. Used in many different genres, for example, to qualitatively assess the content of social networks for politicians and relevant researchers. It is also used to analyze scientific articles, product reviews, online media publications, legal documents, Internet debates and dialog domains.

Natural language processing through higher-level NLP-applications μ_2 simulate intelligent behavior and obvious understanding of natural language and are currently generally divided into the following classes:

7. Text Summarization μ_2^1 (Automatic Summarization) is generation of a readable digest/annotation as a summary in the form of a text fragment from the general analyzed text (for example, a scientific article, publication in a newspaper or magazine).
8. Grammatical Error Correction μ_2^2 is identification and correction of grammatical/stylistic errors at all levels of linguistic analysis (phonology/orthography, morphology, vocabulary, syntax, semantics, pragmatics).
9. Machine Translation μ_2^3 is automatic translation of text from one human language to another using all levels of linguistic analysis, especially grammar, semantics and facts about the real world, etc., based on the solution of AI-complete (AI-hard) tasks of abstract content translation.
10. Dialogue Management μ_2^4 (Dialogue System, Conversational Agent, CA) is the organization of communication between a program and a person, other than chatbots, using text, speech, graphics, tactility, gestures, etc. for two-way communication.
11. Question Answering systems μ_2^5 is determination of the correct answer based on the analysis of a typical human question (for example, "What is the capital of Ukraine?") and an open/complex question (for example, "In what the meaning of existence?").
12. Natural Language Generation μ_2^6 (NLG) is conversion of content from databases/databases or semantic intentions into readable specific human language through an algorithm:
    a. Determination of the content of μ_2^{61} (what information to mention in the text).
    b. Document structuring μ_2^{62} (content transfer template).
    c. Aggregation μ_2^{63} (combining similar sentences to improve the readability and naturalness of the text content).
    d. Lexical selection μ_2^{64} (adding words to concepts).
    e. Generation of reference expressions μ_2^{65} (generation of expressions identifying objects and regions). This task also involves making decisions about pronouns and other types of anaphora.
    f. Implementation of μ_2^{66} (text generation taking into account the rules of syntax, morphology and spelling, for example, the use of verbs in the necessary tenses).
    g. When necessary, use μ_2^{67} machine learning methods (most often LSTM) on a large set of input data and corresponding (human-written) output texts, for example to generate text captions for images.
13. Natural Language Understanding μ_2^7 (NLU) is transformation of text into logical structures for controlling NLP programs, i.e. identification of semantics from a set of possible notations in the form of natural language concepts. The introduction and creation of speech meta-model and ontology is efficient and empirical. To build a formalization of semantics, explicit formalization is used in contrast to implicit assumptions, for example, about a Closed-World Assumption (CWA - any statement

whose truth is not known is false) against an open one (Open-World Assumption OWA – the truth of a statement does not depend on the observer's knowledge of its truth), or subjective YES/NO versus objective TRUE/FALSE.

14. Book Generation μ_2^8 is creation of full-fledged books based on NLG technology μ_2^6, production/associative rules, neural network, factual knowledge and generalization of text μ_2^1.
15. Document AI μ_2^9 is a platform for training an agent to extract specific necessary data from various types of documents. Designed for users without AI, ML and NLP experience to quickly access the necessary content hidden in texts, for example, for lawyers, business analysts and accountants.

4.2. Examples of modeling the processes of solving typical NLP problems

4.2.1. A formal CLS model for identifying viral news headlines

The CLS model for identifying viral news headlines is presented as:

$$S_{vh} = \lambda_1^6 \circ \psi_{NN} \circ \kappa_1 \circ \eta_n \circ \beta_3 \circ \lambda_1^4 \circ \gamma_4, \quad (141)$$

where γ_4 is tokenization; λ_1^4 is recognition of named entities; β_3 is marking of parts of speech; η_n is Ngrams (sequences of elements and their frequencies); κ_1 is clustering; ψ_{NN} is ML based on Neural Networks (NN); λ_1^6 is application of SentiWordNet (a lexical-semantic thesaurus for the analysis of text tonality).

4.2.2. Correction of grammatical and stylistic errors

Another relevant NLP technology is error correction, the main tasks of which are error identification, error correction, and user training. Grammatical error correction μ_2^2 is one of the sub-processes of correcting various types of errors. Bug fixes provided:

$$S_{ec} = \mu_2^2(X), \quad (142)$$

where S_{ec} is the result of identifying and correcting grammatical/stylistic errors at all levels of linguistic analysis (phonology/orthography, morphology, vocabulary, syntax, semantics, pragmatics); μ_2^2 is the basic error correction process in the text data array X.

The detailed error correction process is given as:

$$S_{ec} = \mu_{24}^2 \circ \mu_{23}^2 \circ \mu_{22}^2 \circ \mu_{21}^2, \quad (143)$$

where $X' = \mu_{21}^2(X, R_\mu, R_e, D_l, D_g, D_\beta, \beta_3, \delta_3, \mu_1^1)$ is a pattern matching check based on complex multilayer NLP- rules R_μ based on regular expressions R_e, lexical vand grammatical victionaries, POS-tags D_β and part-of-speech marking β_3, parsing trees δ_3 and coreference resolution μ_1^1;

$X'' = \mu_{22}^2(X', D_t, D_\beta, D_d, \eta_n, \mu_{22}^{21}, \mu_{22}^{22})$ statistical methods for refining corrections (η_n is– Ngrams analysis on based on the set of tokens D_t, POS-tags D_β, and the history of analogues

D_d; μ_{22}^{21} is error identification through the usage statistics with similar words/errors in the text; μ_{22}^{22} is correction to the most likely variant of possible analogues);

$X''' = \mu_{23}^{2}(X'', D_c, R_\eta, D_a, \eta_n, \mu_{23}^{21}, \mu_{23}^{22}, \mu_{23}^{23}, \mu_{23}^{24}, \mu_{23}^{25})$ is machine learning based on a set of classifiers D_c for a specific language, analysis N-grams η_n (for example, bigrams with appropriate analysis of left and right context followed by replacement of the best option from POS Ngrams of the model), machine learning rules R_η as a choice from several correct options or identification correct but rare application, ML-based error detection μ_{23}^{21} and error correction μ_{23}^{22} processes, data annotation D_a for training μ_{23}^{23}, feature selection for training μ_{23}^{24} and classifier training μ_{23}^{25} using random method forest, logistic regression or other (sometimes pattern matching and simple statistical data cannot generalize decision options, for example, identification and selection of a preposition or article in English, and an adjective in Ukrainian; derivational word-formation morphology; run-on sentences, i.e. independent or subordinate clauses are not joined by a conjunction or punctuation);

$S_{ec} = \mu_{24}^{2}(X''', D_c, R_\eta, D_a, \eta_n, \mu_{24}^{21}, \mu_{24}^{22})$ is neural machine translation based on Noisy channel translation processes μ_{24}^{21} (spelling check, question answers, speech recognition and machine translation based on finding a predicted word with a given word, where symbols are somehow encrypted) and Round-trip translation μ_{24}^{22} (two-way translation from the source language to the target language to assess the quality/accuracy of the result).

The N-gram model assigns probabilities to sentences and sequences of words based on counting N-grams, for example, according to the Markov assumption:

$$p(x_j^n) \approx \prod_{i=1}^{n} p(x_i | x_{i-1}), \tag{144}$$

where x_j^n is the j-th chain or sentence of n words; x_i is the current word in the j-th chain or sentence; x_{i-1} is the previous word in the j-th chain or sentence. To identify an error, it is necessary to determine grammatical or stylistic features using the rules of marking parts of speech β_3, parsing: dependencies δ_3^1 and constituents δ_3^2 or in the form of some combination of these methods δ_3^3.

5. Conclusions

The general structure of CLS processing of textual content in the Ukrainian language and the conceptual scheme/model of the functioning of a typical CLS based on modeling the interaction of the main processes and IS components have been developed.

Modeling of the main NLP processes of CLS was carried out due to the interaction of the main processes/components of IS and methods of linguistic processing of text content adapted to the Ukrainian language on the basis of grapheme, morphological, lexical, syntactic, semantic, structural, ontological and pragmatic analysis, which allowed to improve the IT of intellectual analysis of the text flow for solving a specific NLP problem. This ensured the adaptation of NLP processes for the analysis of Ukrainian-language textual content. A formal model of a computer linguistic system for processing Ukrainian-language textual content was developed and described, which made it possible to determine the main

structural elements and operators of natural language processing at each level of text analysis such as grapheme/phonological, morphological, syntactic, semantic, referential, structural, ontological and pragmatic. Due to the complexity of the morphology of the Ukrainian language, detailed attention is paid to the description of the model of morphological analysis of textual content. Examples of modeling processes for solving typical NLP problems such as CLS identification of viral news headlines and correction of grammatical and stylistic errors are given.

References

[1] N. Khairova, T. Hamon, N. Grabar, Y. Burov, Preface: Computational Linguistics Workshop, CEUR Workshop Proceedings 3396 (2023). URL: https://ceur-ws.org/Vol-3396/.

[2] V. Lytvyn, N. Sharonova, I. Jonek-Kowalska, A. Kowalska-Styczen, V. Vysotska, Y. Kupriianov, O. Kanishcheva, O. Cherednichenko, T. Hamon, N. Grabar, Preface: Computational Linguistics and Intelligent Systems, CEUR Workshop Proceedings 3171 (2022). URL: https://ceur-ws.org/Vol-3171/.

[3] N. Sharonova, V. Lytvyn, O. Cherednichenko, Y. Kupriianov, O. Kanishcheva, T. Hamon, N. Grabar, V. Vysotska, A. Kowalska-Styczen, I. Jonek-Kowalska, Preface: Computational Linguistics and Intelligent Systems, CEUR Workshop Proceedings 2870 (2021). URL: https://ceur-ws.org/Vol-2870/.

[4] V. Lytvyn, V. Vysotska, T. Hamon, N. Grabar, N. Sharonova, O. Cherednichenko, O. Kanishcheva, Preface: Computational Linguistics and Intelligent Systems, CEUR Workshop Proceedings 2604 (2020). URL: https://ceur-ws.org/Vol-2604/.

[5] V. Lytvyn, N. Sharonova, T. Hamon, O. Cherednichenko, N. Grabar, A. Kowalska-Styczen, V. Vysotska, Preface: Computational Linguistics and Intelligent Systems, CEUR Workshop Proceedings 2362 (2019). URL: https://ceur-ws.org/Vol-2362/.

[6] S. Hlibko, N. Vnukova, D. Davydenko, V. Pyvovarov, V. Avanesian, The Use of Linguistic Methods of Text Processing for the Individualization of the Bank's Financial Servise, CEUR Workshop Proceedings 3403 (2023) 157-167.

[7] N. Vnukova, D. Davydenko, S. Hlibko, V. Shorokh, Mastering Computer Linguistics for the Designation of Risks in Cooperation Communications, CEUR Workshop Proceedings 3171 (2022) 376-386.

[8] L. Savytska, N. Vnukova, I. Bezugla, V. Pyvovarov, M. Turgut Sübay, Using Word2vec Technique to Determine Semantic and Morphologic Similarity in Embedded Words of the Ukrainian Language, CEUR Workshop Proceedings 2870 (2021) 235-248.

[9] T. Batiuk, V. Vysotska, R. Holoshchuk, S. Holoshchuk, Intelligent System for Socialization of Individual's with Shared Interests based on NLP, Machine Learning and SEO Technologies, CEUR Workshop Proceedings 3171 (2022) 572-631.

[10] A. Mykytiuk, V. Vysotska, S. Albota, Spam Filtration System with the Use of Machine Learning Technology, in: Proceedings of the IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), 22-25 Sept., Lviv, Ukraine. 2021, vol. 1, pp. 124–130.

[11] N. Liubchenko, A. Podorozhniak, V. Oliinyk, Research Application of the Spam Filtering and Spammer Detection Algorithms on Social Media, CEUR Workshop Proceedings 3171 (2022) 116-126.
[12] V.B. Fernandes, et al., Corrigendum to "A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification", Applied Soft Computing Journal 55 (2017) 565.
[13] V. Basto-Fernandes, et al., A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification, Applied Soft Computing Journal 48 (2016) 111–123.
[14] I Yevseyeva, V. Basto-Fernandes, D. Ruano-Ordás, J.R. Méndez, Optimising anti-spam filters with evolutionary algorithms, Expert Systems with Applications 40(10) (2013) 4010–4021.
[15] D. Ruano-Ordás, V. Basto-Fernandes, I. Yevseyeva, J.R. Méndez, Evolutionary multi-objective scheduling for anti-spam filtering throughput optimization, Lecture Notes in Computer Science 10334 (2017) 137–148.
[16] O. Cherednichenko, O. Kanishcheva, Readability Evaluation for Ukrainian Medicine Corpus(UKRMED), CEUR Workshop Proceedings 2870 (2021) 402-412.
[17] V. Bobicev, O. Kanishcheva, O. Cherednichenko, Sentiment Analysis in the Ukrainian and Russian News, in: Procedeengs of the First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017, pp. 1050-1055.
[18] O. Cherednichenko, N. Babkova, O. Kanishcheva, Complex Term Identification for Ukrainian Medical Texts, CEUR Workshop Proceedings 2255 (2018) 146-154.
[19] I. Balush, V. Vysotska, S. Albota, Recommendation System Development Based on Intelligent Search, NLP and Machine Learning Methods, CEUR Workshop Proceedings 2917 (2021) 584-617.
[20] V. Vysotska, S. Mazepa, L. Chyrun, O. Brodyak, I. Shakleina, V. Schuchmann, NLP Tool for Extracting Relevant Information from Criminal Reports or Fakes/Propaganda Content, in: Procedeengs of the IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, pp. 93-98, doi: 10.1109/CSIT56902.2022.10000563.
[21] O. Prokipchuk, V. Vysotska, P. Pukach, V. Lytvyn, D., Uhryn, Y. Ushenko, Z. Hu, Intelligent Analysis of Ukrainian-language Tweets for Public Opinion Research based on NLP Methods and Machine Learning Technology, International Journal of Modern Education and Computer Science 15(3) (2023) 70–93.
[22] V. Husak, O. Lozynska, I. Karpov, I. Peleshchak, S. Chyrun, A. Vysotskyi, Information System for Recommendation List Formation of Clothes Style Image Selection According to User's Needs Based on NLP and Chatbots, CEUR workshop proceedings 2604 (2020) 788-818.
[23] J. Deriviere, T. Hamon, A. Nazarenko, A scalable and distributed NLP architecture for web document annotation, Lecture Notes in Computer Science 4139 (2006) 56–67.
[24] M. Boyè, T.M. Tran, N. Grabar, NLP-oriented contrastive study of linguistic productions of alzheimer's and control people, Lecture Notes in Computer Science 8686 (2014) 412–424.

[25] A. Berko, V. Andrunyk, L. Chyrun, M. Sorokovskyy, O. Oborska, O. Oryshchyn, M. Luchkevych, O. Brodovska, The Content Analysis Method for the Information Resources Formation in Electronic Content Commerce Systems, CEUR Workshop Proceedings 2870 (2021) 1632-1651.
[26] R. Bekesh, L. Chyrun, P. Kravets, A. Demchuk, Y. Matseliukh, T. Batiuk, I. Peleshchak, R. Bigun, I. Maiba, Structural modeling of technical text analysis and synthesis processes, CEUR Workshop Proceedings 2604 (2020) 562–589.
[27] O. Artemenko, V. Pasichnyk, N. Kunanets, K. Shunevych, Using sentiment text analysis of user reviews in social media for e-tourism mobile recommender systems, CEUR workshop proceedings 2604 (2020) 259-271.
[28] N. Grabar, T. Hamon, Automatic Detection of Temporal Information in Ukrainian General-language Texts, CEUR Workshop Proceedings 2136 (2018) 1-10.
[29] V. Shyrokov, "Accuracy" vs "Unambiguity" in Linguistics, CEUR Workshop Proceedings 2870 (2021) 1-5.
[30] O. Kozlov, Structural Optimization of Fuzzy Systems based on Determination of Linguistic Terms Number, CEUR Workshop Proceedings 2917 (2021) 174-185.
[31] O. Yurchenko, N. Ugolnikova, Linguistic Methods in Social Media Marketing, CEUR Workshop Proceedings 2870 (2021) 743-754.
[32] E. Fedorov, O. Nechyporenko, Linguistic Constructions Translation Method Based on Neural Networks, CEUR Workshop Proceedings 3396 (2023) 295-306.
[33] E. Fedorov, O. Nechyporenko, Dynamic Stock Buffer Management Method Based on Linguistic Constructions, CEUR Workshop Proceedings 2870 (2021) 1742-1753.
[34] A. Taran, The Role of Keyword Language in the Database of World Slavic linguistics "iSybislaw", CEUR Workshop Proceedings 3171 (2022) 266-276.
[35] A. Taran, Terminology of Computational Linguistics in Terms of Indexing and Information Retrieval in the System "iSybislaw", CEUR Workshop Proceedings 2870 (2021) 225-234.
[36] A. Taran, Information-retrieval System "Base of the World Slavic Linguistics (iSybislaw)" in Language Education, CEUR workshop proceedings 2604 (2020) 590-599.
[37] I. Bekhta, N. Hrytsiv, Computational Linguistics Tools in Mapping Emotional Dislocation of Translated Fiction, CEUR Workshop Proceedings 2870 (2021) 685-699.
[38] D. Sitnikov, P. Sytnikova, A. Kovalenko, Methods of Eliminating Features from Linguistic Equations, CEUR Workshop Proceedings 2870 (2021) 877-889.
[39] S. Albota, Linguistically Manipulative, Disputable, Semantic Nature of the Community Reddit Feed Post, CEUR Workshop Proceedings 2870 (2021) 769-783.
[40] S. Albota Resolving conflict situations in reddit community driven discussion platform. CEUR Workshop Proceedings 2604 (2020) 215–226.
[41] S. Albota, Linguistic and Psychological Features of the Reddit News Post, in: Procedenngs of the IEEE 15th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2020, 1, pp. 295–299.
[42] S. Albota, A. Peleshchyshyn, Contradictory statement as a basis for conflict resolution strategies, CEUR Workshop Proceedings 2588 (2020) 336–345.

[43] S. Albota, A. Peleshchyshyn, D. Dosyn, Linguistic traces of psychological manipulations in discussions of wikipedia talk pages, CEUR Workshop Proceedings 2386 (2019) 183–193.
[44] S. Albota, V. Yakovyna, A. Peleshchyshyn, Discussions of Wikipedia Talk Pages: manipulations detected by lingual-psychological analysis, CEUR Workshop Proceedings 2392 (2019) 309–320.
[45] S. Albota, A. Peleshchyshyn, Requirements for the Linguistic Quality Control of Wikipedia Article, in: Procedeengs of the IEEE 14th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2019, 3, pp. 16–19.
[46] S. Albota, A. Peleshchyshyn, O. Markovets, V. Vus, A formal approach to modeling the characteristics of users of social networks regarding information security issues, Advances in Intelligent Systems and Computing 902 (2020) 485–494.
[47] S Albota, N. Melnykova, M. Buchyn, S. Fedushko S. Kashuba The special ways for processing personalized data during voting in elections, Advances in Intelligent Systems and Computing 1080 (2020) 781–791.
[48] A. Peleshchyshyn, O. Markovets, V. Vus, S. Albota, Identifying specific roles of users of social networks and their influence methods, in: Procedeengs of the IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2, pp. 39–42.
[49] S. Albota, Linguistically manipulative, disputable, semantic nature of the community reddit feed post, CEUR Workshop Proceedings 2870 (2021) 769–783.
[50] O. Markovets, S. Albota, O. Horpyniuk, Information consolidation on users of social networks to determine their credibility, CEUR Workshop Proceedings 2870 (2021) 784–798.
[51] S. Bhatia, M. Sharma, K. K. Bhatia, P. Das, Opinion target extraction with sentiment analysis, International Journal of Computing 17(3) (2018) 136-142. doi:10.47839/ijc.17.3.1033.
[52] O. Bisikalo, V. Vysotska, Linguistic analysis method of Ukrainian commercial textual content for data mining, CEUR Workshop Proceedings 2608 (2020) 224-244.
[53] V. Vysotska, Ukrainian Participles Formation by the Generative Grammars Use, CEUR workshop proceedings 2604 (2020) 407-427.
[54] V. Vysotska, S. Holoshchuk, R. Holoshchuk, A comparative analysis for English and Ukrainian texts processing based on semantics and syntax approach, CEUR Workshop Proceedings 2870 (2021) 311-356.
[55] K. Tymoshenko, V. Vysotska, O. Kovtun, R. Holoshchuk, S. Holoshchuk, Real-time Ukrainian text recognition and voicing, CEUR Workshop Proceedings 2870 (2021) 357-387.
[56] N. Darchuk, K. Sergii, V. Sorokin, Logico-Linguistic Model of Ukrainian Text, CEUR Workshop Proceedings 3396 (2023) 20-31.
[57] N. Darchuk, V. Sorokin, Parameterization of the Ukrainian Text Corpus Based on Parsing Results, CEUR Workshop Proceedings 3171 (2022) 256-265.

[58] N. Hrytsiv, I. Bekhta, M. Tkachivska, V. Byalyk, Sylvia Plath's I felt-Narrative Label of The Bell Jar in Ukrainian Translation: Tagging Textness Features, CEUR Workshop Proceedings 3171 (2022) 240-255.
[59] Z. Kunch, O. Lytvyn, I. Mentynska, Modern Ukrainian Electronic Dictionaries: the Problem of Implementing Spelling Changes, CEUR Workshop Proceedings 3396 (2023) 32-47.
[60] O. Levchenko, M. Dilai, Key Colour Terms in the Ukrainian Prose Fiction of the 21st Century, CEUR Workshop Proceedings 3171 (2022) 49-60.
[61] O. Levchenko, N. Romanyshyn, D. Dosyn, Method of Automated Identification of Metaphoric Meaning in Adjective + Noun Word Combinations (Based on the Ukrainian Language), CEUR Workshop Proceedings 2386 (2019) 370-380.
[62] V. Starko, A. Rysin, VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool, CEUR Workshop Proceedings 3171 (2022) 61-70.
[63] V. Starko, Implementing Semantic Annotation in a Ukrainian Corpus, CEUR Workshop Proceedings 2870 (2021) 435-447.
[64] O. Synchak, V. Starko, Ukrainian Feminine Personal Nouns in Online Dictionaries and Corpora, CEUR Workshop Proceedings 3171 (2022) 775-790.
[65] N. Romanyshyn, Application of Corpus Technologies in Conceptual Studies (based on the Concept Ukraine Actualization in English and Ukrainian Political Media Discourse), CEUR workshop proceedings 2604 (2020) 472-488.
[66] M. Shvedova, The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality, CEUR workshop proceedings 2604 (2020) 489-506.
[67] M. Shvedova, N. Prydvorova, I. Skibina, Normalization of Early Modern Ukrainian in GRAC: the Case of Lesia Ukrainka's Works, CEUR Workshop Proceedings 3171 (2022) 71-80.
[68] S. Kubinska, R. Holoshchuk, S. Holoshchuk, L. Chyrun, Ukrainian Language Chatbot for Sentiment Analysis and User Interests Recognition based on Data Mining, CEUR Workshop Proceedings 3171 (2022) 315-327.
[69] N. Borysova, K. Melnyk, N. Babkova, Z. Kochuieva, V. Melnyk, Gender Classification of Surnames: Ukrainian aspect, CEUR Workshop Proceedings 3171 (2022) 354-364.
[70] A. Dmytriv, S. Holoshchuk, L. Chyrun, R. Holoshchuk, Comparative Analysis of Using Different Parts of Speech in the Ukrainian Texts Based on Stylistic Approach, CEUR Workshop Proceedings 3171 (2022) 546-560.
[71] K. S. Mandziy, U. V. Yurlova, M. P. Dilai, English-Ukrainian Parallel Corpus of IT Texts: Application in Translation Studies, CEUR Workshop Proceedings 3171 (2022) 724-736.
[72] K. Vyrodov, A. Chupryna, R. Kotelnykov, Detecting of Anti-Ukrainian Trolling Tweets, CEUR Workshop Proceedings 3396 (2023) 48-62.
[73] L. Kobylyukh, Z. Rybchak, O. Basystiuk, Analyzing the Accuracy of Speech-to-Text APIs in Transcribing the Ukrainian Language, CEUR Workshop Proceedings 3396 (2023) 217-227.

[74] K. Datsyshyn, Z. Haladzhun, N. Kunanets, O. Hotsur, N. Veretennikova, Neologisms with the Prefix Anti- in the Ukrainian Online Media in the Covid-19 Pandemic Period, CEUR Workshop Proceedings 3171 (2022) 192-211.
[75] Z. Haladzhun, K. Datsyshyn, Y. Bidzilya, N. Kunanets, N. Veretennikova, "Anti-vaccinationists&Anti-vax": Linguistic Means of Actualizing Assessment in the Headlines and Leads of Ukrainian Text Media, CEUR Workshop Proceedings 3396 (2023) 118-129.
[76] Y. Chychkarov, O. Zinchenko, Handwritten Ukrainian Character Recognition using a Convolutional Neural Networks and Synthetic Dataset, CEUR Workshop Proceedings 3426 (2023) 109-121.
[77] T. Basyuk, A. Vasyliuk, Peculiarities of an Information System Development for Studying Ukrainian Language and Carrying out an Emotional and Content Analysis, CEUR Workshop Proceedings 3396 (2023) 279-294.
[78] A. Dmytriv, V. Vysotska, M. Bublyk, The Speech Parts Identification for Ukrainian Words Based on VESUM and Horokh Using, in: Proceedings of the IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), 22-25 Sept., Lviv, Ukraine, 2021, vol. 2, pp. 21–33.
[79] V. Vysotska, O. Markiv, S. Teslia, Y. Romanova, I. Pihulechko, Correlation Analysis of Text Author Identification Results Based on N-Grams Frequency Distribution in Ukrainian Scientific and Technical Articles, CEUR Workshop Proceedings 3171 (2022) 277-314.
[80] V. Lytvyn, P. Pukach, V. Vysotska, M. Vovk, N. Kholodna, Identification and Correction of Grammatical Errors in Ukrainian Texts Based on Machine Learning Technology, Mathematics 11 (2023) 904. doi:10.3390/math11040904.
[81] P. Kryndach, V. Vysotska, S. Chyrun, L. Chyrun, S. Goloshchuk, R. Holoshchuk, Analysis of Semantic Relationships in Ukrainian Text Content Based on Word2Vec and Machine Learning, in: Proceedings of the IEEE 18th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, 19-21 October, 2023.
[82] B. Rusyn, V. Vysotska, L. Pohreliuk, Model and architecture for virtual library information system, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 37-41. doi: 10.1109/STC-CSIT.2018.8526679.
[83] B. Rusyn, V. Lytvyn, V. Vysotska, M. Emmerich, L. Pohreliuk, The Virtual Library System Design and Development, Advances in Intelligent Systems and Computing 871 (2019) 328-349. doi: 10.1007/978-3-030-01069-0_24.
[84] L. Chyrun, Y. Burov, B. Rusyn, L. Pohreliuk, O. Oleshek, A. Gozhyj, I. Bobyk, Web resource changes monitoring system development, CEUR Workshop Proceedings 2386 (2019) 255–273.
[85] B. Rusyn, L. Pohreliuk, A. Rzheuskyi, R. Kubik, Y. Ryshkovets, L. Chyrun, S. Chyrun, A. Vysotskyi, V.B. Fernandes, The mobile application development based on online music library for socializing in the world of bard songs and scouts' bonfires, Advances in Intelligent Systems and Computing 1080 (2020) 734–756. doi**:** 10.1007/978-3-030-33695-0_49.

[86] A. Demchuk, B. Rusyn, L. Pohreliuk, A. Gozhyj, I. Kalinina, L. Chyrun, N. Antonyuk, Commercial content distribution system based on neural network and machine learning, CEUR Workshop Proceedings 2516 (2019) 40–57.
[87] B. Rusyn, L. Pohreliuk, O. Kapshii, J. Varetskyy, A. Demchuk, I. Karpov, A. Gozhyj, V. Gozhyj, I. Kalinina, An Intelligent System for Commercial of Information Products Distribution Based SEO and Sitecore CMS, CEUR workshop proceedings 2604 (2020) 760-777.
[88] V. Andrunyk, A. Berko, B. Rusyn, L. Pohreliuk, S. Chyrun, B. Dokhniak, I. Karpov, M. Krylyshyn, Information System of Photostock Web Galleries Based on Machine Learning Technology, CEUR workshop proceedings 2604 (2020) 1032-1059.
[89] L. Chyrun, Y. Burov, B. Rusyn, L. Pohreliuk, O. Oleshek, A. Gozhyj, I. Bobyk, Web Resource Changes Monitoring System Development, CEUR Workshop Proceedings 2386 (2019) 255-273.
[90] O. Mediakov, V. Vysotska, D. Uhryn, Y. Ushenko, C. Hu, Information technology for generating lyrics for song extensions based on transformers, International Journal of Modern Education and Computer Science 16(1) (2024) 23–36.
[91] O. Voloshynskyi, V. Vysotska, R. Holoshchuk, S. Goloshchuk, S. Chyrun, D. Zahorodnia. Sign Language Digits Recognition Technology Based on a Convolutional Neural Network, in: Proceedings of the IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Dortmund, Germany, 07-09 September 2023, pp. 27–32.
[92] V. Turchenko, E. Chalmers, A. Luczak, A deep convolutional auto-encoder with pooling – unpooling layers in caffe, International Journal of Computing 18(1) (2019) 8-31. doi:10.47839/ijc.18.1.1270.
[93] S. Anfilets, S. Bezobrazov, V. Golovko, A. Sachenko, M. Komar, R. Dolny, V. Kasyanik, P. Bykovyy, E. Mikhno, O. Osolinskyi, Deep multilayer neural network for predicting the winner of football matches, International Journal of Computing 19(1) (2020) 70-77. doi:10.47839/ijc.19.1.1695.