

Modern State and Prospects of Information Technologies Development for Natural Language Content Processing

Victoria Vysotska

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

Abstract

The study aims to develop a new method of building computer linguistic systems (CLS) for processing Ukrainian-language text content for solving various NLP problems based on the application of intellectual analysis of text flow from information resources. This became possible thanks to the combination of linguistic analysis methods adapted to the Ukrainian language, improved information technology for processing information resources, machine learning technology and a set of metrics for evaluating the effectiveness of the functioning of computer linguistic systems. The peculiarity of such CLS construction is based on the basic principle of system modularity (presence/absence of basic and additional modules), which facilitates the development of specific modules according to the requirements for the relevant processes implementation for solving a specific NLP problem. The developed methods and tools made it possible to build computer linguistic systems for processing Ukrainian-language text content to solve a specific NLP problem according to the needs of the permanent/potential target audience based on the analysis of the history of their actions on the CLS Web resource. The improved technology of intellectual processing of Ukrainian-language textual content, unlike the existing ones, supports the modularity principle of the typical CLS architecture for solving a specific NLP problem and analysing a set of parameters and metrics of system performance by the behaviour of the target audience. This made it possible to develop a general typical CLS structure and a conceptual diagram/model of the operation of a typical CLS based on the modelling of the interaction of the main processes and components. Improvement of methods of processing information sources (resources), such as integration, management and support of Ukrainian-language content, made it possible to adapt the process of intellectual analysis of text flow to solving various NLP tasks.

Keywords

Computer linguistic system, intelligent search system, NLP, Ukrainian language, information resource, system performance metrics, machine learning, target audience


1. Introduction


The well-known expression of the English banker, businessman and financier Nathan Rothschild "He who owns information owns the world" is very relevant today, which has been in use for over two centuries. The modern century is an era of information technology (IT) and artificial intelligence, which surround the average person everywhere in everyday life. And where there is a person, there is a natural language. Therefore, the combination of information technologies, artificial intelligence and natural language processing is relevant in human society for solving everyday tasks. The solution of such tasks is entrusted to a modern young scientific direction such as computer linguistics. The difficulty lies not only in solving non-typical NLP problems but in adapting or creating new models, methods and technologies for processing a specific natural language. Each natural language is unique, with its flavour of rules, history, grammar, exceptions, and features of generating linguistic units to convey meaning. On average, a person studies for 10-15 years to understand everyday life, another 10-15 years to adapt to a profession, and the language itself and its depth can be studied and explored for a lifetime. There is no such luxury as time to automate the processes of working out a specific natural language. In addition, the limited financing of similar projects or their absence at all, competition with well-known companies and

COLINS-2024: 8th International Conference on Computational Linguistics and Intelligent Systems, April 12-13, 2024, Lviv, Ukraine

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

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

 © 2024 Copyright for this paper by its authors.

 Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

the presence on the market of their developed commercial projects significantly reduce the motivation of scientists' work in this direction. Usually, each successful project for the development of computer linguistic systems (CLS) is designed for a specific task and at the same time is one-time and closed (for example, Facebook, Siri, Google Assistant, Amazon Alexa, Microsoft Cortana, Bixby, Voice Mate, Alisa, Abby Lingvo, Microsoft Word, Grammarly, Google Translation, PROMT, CuneiForm, Trados, OmegaT, Wordfast, Dragon, IBM via voice, Speereo, Finereader, Tesseract, OCRopus, etc.) without being able to read the content to willing IT professionals/specialists. There are quite rare cases when such projects are given open access and an opportunity to get acquainted with their structure and other content.

The Internet, mobile applications, information systems, and social networks – bottomless sources of information are constantly present around us. On the one hand, it helps to solve many everyday and professional tasks, but on the other hand, it complicates the life process due to the need to navigate in this chaos of information space. In addition, it is a source of manipulation of people's consciousness through propaganda, fakes both in everyday life (for example, through advertising), and in information warfare, etc.

Nowadays, much online information is subject to regional censorship in certain territorial regions due to political, economic, social, religious and other factors, such as to control or manage the opinion of the people of that region. The reasons can be various factors. At the same time, fake information is spread both purposefully and randomly/chaotically in the Internet environment. It is easy for an average person to get lost and navigate in this mass of content flow with opposing facts and causes of events/phenomena. It is unethical, illegal and impractical to control exactly what to show or hide (to censor content) among Internet content to the average user in democratic states. This is one of the first steps in the transition to totalitarianism. But providing information, for example, to journalists about a possible thematic fake for conducting a journalistic investigation or warning the average reader about the possibility of disinformation in this content/resource is, on the one hand, support for freedom of speech, and on the other hand, giving a person the opportunity to choose what to believe and what not to believe. At the same time, it provides an opportunity to gain an understanding of events and orientation in a large flow of information both for solving everyday tasks and adjusting business strategies, etc. Significant and massive dissemination of (dis)information against the background of the war in Ukraine without appropriate analysis potentially leads to panic among the relevant stratum/region of the population, significantly affecting the process of adjusting plans/strategies of business, social services, etc. Against the background of the information war, a lot of time and resources are spent on the appropriate collection, analysis and formation of appropriate conclusions regarding the content of the relevant content. This is also influenced by the language of the information, which may partially/significantly change the content when translated. KLS will not be able to completely replace human activity in this direction. However, it can be a significant helper for quickly forming relevant bases of such content and reacting to local changes or the dynamics of changes in the content flow, marking certain content as potentially fake in a certain percentage. The difficulty lies in the language of the content itself. In comparison with English-language content, Ukrainian/russian languages are quite difficult to automatically process, especially the extraction and analysis of semantics. Today, there are many computer linguistic systems for various purposes, even for processing Ukrainian-language textual content. But these are usually commercial projects of a closed type (there are no publications or access to the administrative part) and most often they are foreign projects. There seem to be a lot of publications to understand how the natural language processing process generally works, especially for English texts. However, applying these models, methods, algorithms and technologies directly to Ukrainian-language textual content does not lead to almost any positive result. Already at the level of morphological analysis, a significant conflict arises between the developed methods and the incoming Ukrainian text - the output is not correct. For example, for a simple Porter algorithm (stemming) without a corresponding modification, it will not be correct to separate the base of the word from the inflexion, which will lead to incorrect identification of the keywords of the texts, which in turn affects any NLP task where it is necessary to quickly identify a set of keywords (rubrication, search, annotation, etc.). Determining the main processes

and features of the linguistic analysis of Ukrainian-language texts will greatly facilitate the stages of processing the text flow of content such as integration, support and content management. In turn, the adaptation of the processes of intellectual analysis of text content with the identification of functional requirements for the corresponding modules of the CLS will lead to the possibility of developing a typical architecture of such systems based on the principle of modularity (adding components depending on the content of the NLP task and the purpose of the CLS).

To solve most NLP problems, the words of the relevant textual content are processed, analysed and researched as a result of the work of one or more authors in a specific dialect of a certain language (the best measure of the variation of the author's speech characteristics), of a certain style (dialogue/monologue) and genre (an auxiliary measure of variation features of the author's speech) at a certain time, in a certain place, for a certain purpose/function. There are more than 7 thousand languages in the modern world. NLP algorithms are most useful when they are applied to many languages. Most NLP tools are usually developed for the official languages of large industrialized countries (English, Chinese, German, Russian, etc.) and this is a very limited range of natural languages (out of a couple of dozen). For most of the world's languages, either no NLP tools are developed, or no significant attention is paid (surface development) or highly specialized commercial projects. But usually, most of the content consists of text in more than one language. Therefore, it is advisable to support the development of NLP tools in several languages according to their purpose, for example, for the classification of text content in the scientific and technical Ukrainian language, it is advisable to use a combination of NLP techniques not only for the analysis of the Ukrainian language but at least English due to the presence of specific terminology and habits speakers to use English analogues from the subject area.

In addition, most natural languages have several regional, social or professional dialects or slang/slang. This makes it possible to maintain appropriate dictionaries not only for content classification but also, for example, for identifying the probable author of the corresponding text. At the same time, some languages are constantly developing and changing at different speeds, which significantly affects the quality of processing new modern content. Simply changing the RE-rules will not solve the problem, as all the old contents of the content will not be rewritten. It is then necessary to introduce the concept of classification of old/new RE-rules, for example, the morphological processing of words and the support of relevant dictionaries.

Speakers/writers quite often use multiple languages (based on automatic code-switching according to the content) in a single communicative written/audio content of the appropriate genre (news, fantasy novel, scientific article or detective story, etc.) and subject matter (technical, medical, social, etc.). The source of the text also affects the processing features, for example, spoken (business meetings, telephone conversations, court proceedings, medical advice, recording of a public speech, etc.) or official documents (laws, orders, etc.). The text reflects the demographic and social characteristics of the author/speaker such as age range, gender, level of education (not only the level of literacy and field of education but the level of depth of knowledge), origin, socio-economic status, etc. Also, the text reflects the approximate period of the appearance of the world due to the peculiarities of the language in different periods - each language changes over time. Since language is so situational when developing computational NLP models, it is important to take into account the characteristics of the author, the context of the text, the purpose of the assignment, etc. Ukrainian-language textual content, regardless of style, usually contains a significant amount of unstructured abstract information. It is a meaningful chain of linguistic units with a predetermined structure, integrity and coherence. Correct, operative and full-fledged content analysis of the relevant Ukrainian-language text allows for solving many modern NLP tasks. Parsing Ukrainian textual content into lexemes based on finite automata and Chomsky grammar is a classic approach. However, it does not solve the main problems of processing Ukrainian-language textual content.

The creation of any NLP application for an arbitrary natural language from more than 7,000 known languages and dialects is based precisely on the researched data (large monolingual/parallel text corpora of more than hundreds of millions of words and linguistic resources) of a specific language. Only about 20 natural languages (English, Chinese, Spanish and other Western European languages, Japanese, etc.) have relevant research results and meet the

requirements for the development of different complexity of CLS. Unfortunately, in modern realities, the Ukrainian language is considered in the international scientific community to be an exotic language with a low resource index, that is, it does not have enough educational, research and processed data for the development of modern NLP applications while meeting the relevant needs of society, in particular, in cyber security (detection of fakes and propaganda, so-called trolls/bots in social networks, etc.), sociology (analysis of the dynamics of changes in public opinion on certain thematic issues, etc.), philology (automatic research of large data sets of different thematic directions and different periods), psychology (analysis of the psychological portrait of a person based on posts in social networks, identification of post-traumatic stress disorder in participants of hostilities or occupation, etc.) and in other important areas of modern Ukraine. The proposed methodology is based on the application of advanced technologies of linguistic analysis and intellectual processing of Ukrainian-language textual content, further training of CLS on reliable data and analysis of the obtained results to find features and regularities of the appearance of linguistic events (keywords, stable phrases, metrics of the author's style, etc.). The above shows the relevance of research in the direction of building computer linguistic systems for processing Ukrainian-language textual content to solve various NLP tasks according to the needs of the target audience of Ukraine and international society.

2. Related works

2.1. The concept of computer linguistic systems

The modern development of IT is at the intersection of globalization and informatization. The rapid rate of growth of the informatization of society is directly related to the rate of development and introduction of IT processing of natural language (Natural-Language Processing, NLP), the main tools of which are computer linguistic systems (CLS). According to [1], there are two different interpretations of the existing term linguistic system (LS):

1. A set of linguistic units of the corresponding speech level (phonology, morphology, syntax, etc.) taking into account their unity and interconnection; types of linguistic units and the rules of their formation, transformation and combination. The identification of language as a language is attributed to F. de Saussure and is based on the works of W. von Humboldt and J. Baudouin de Courtenay.
2. A set of oppositions (facts) at the appropriate linguistic level using metalanguage for description and identification.

In [2], the linguistic information system or LS is defined as a system that an individual uses for his speech activity. According to the interpretation of the Stanford Encyclopaedia [3], computer linguistics (Computational Linguistics, CL) is a scientific and engineering discipline for finding approaches to understanding written and spoken language by computer means, as well as creating methods for processing natural language. Since language is a mirror of the mind, computer language understanding also contributes to the understanding of the thought process and the content of intelligence. If natural language is the most natural and universal means of communication, then appropriate software (software) with linguistic competence should greatly facilitate human interaction through computers with each other and information systems (IS) of all kinds to meet everyday needs, for example in IIS/analysis huge text arrays of data and other Websites. Accordingly, CLS is intended for solving NLP problems according to user needs. The main features of CLS are the use of methods of artificial intelligence (Artificial Intelligence, AI), applied linguistics (Applied Linguistics, AL), system analysis and IT for understanding natural information when solving various NLP tasks both in everyday human life and modern research of a specialized scientific direction (Fig. 1). The main objects of computer linguistics are content - arbitrary semi-structured and partially formalized information, presented orally by speech, written text, visually/emotionally by facial expressions and gestures, graphically by emoticons/images and/or by any other means of transmission. Content is a collection of various types of data (text, sound, service, commercial, additional, etc.), which form a corresponding set

of meta-models (a description of the structure and features of the model's functioning) and template models implemented within a specific information system (ISO/IEC/IEEE 24765:2010(E), 3.1398, ISO/IEC 15474-1:2002, Information technology). There are quite a lot of publications and studies on solving various NLP problems for processing English-language texts. There are significantly fewer studies for Slavic languages, in particular, for Ukrainian. In general, there are no publications on development recommendations, functional requirements, general structure and typical architecture of CLS.

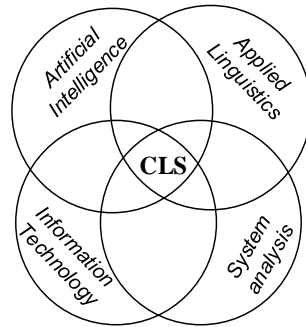


Figure 1: The main modern directions for the synthesis of CLS

Typically, each successful CLS development project is task-specific and both disposable and closed (e.g. Siri, Amazon Alexa, Google Assistant, Grammarly, Abby Lingvo, Facebook, Voice Mate, Bixby, Microsoft Cortana, Microsoft Word, Google Translation, PROMT, CuneiForm, Trados, OmegaT, Wordfast, Dragon, IBM via voice, Speereo, Finereader, Tesseract, OCRopus, etc.) without the possibility of familiarizing the content to willing IT professionals/specialists. There are quite rare cases when such projects are given open access and an opportunity to get acquainted with their structure and other content. Accordingly, research in the direction of analysis and synthesis of CLS, in particular, for the processing of Ukrainian-language texts for today is relevant and promising [4]-[9], for example, on <https://victana.lviv.ua/>.

2.2. General classification of computer linguistic systems

Today, the field of CL is developing rapidly, but most projects are commercialized and disposable. Therefore, there is no single unequivocal approach, typical general recommendations, advice and requirements for the design, analysis, development and synthesis of relevant CLS. There is also no consensus on the typification, categorization and classification of CLS. This makes it much more difficult to navigate in the chaos of publications and research, which methods and tools need to be applied to effectively obtain the desired results, in particular, when solving a specific NLP task of processing Ukrainian-language texts [7]-[9]. For example, according to the classification by Grammarly, there are only three main types of CLS: analytical, transformational and combined (Fig. 2). There are many more types and possibilities of CLS than described in [10]-[15].

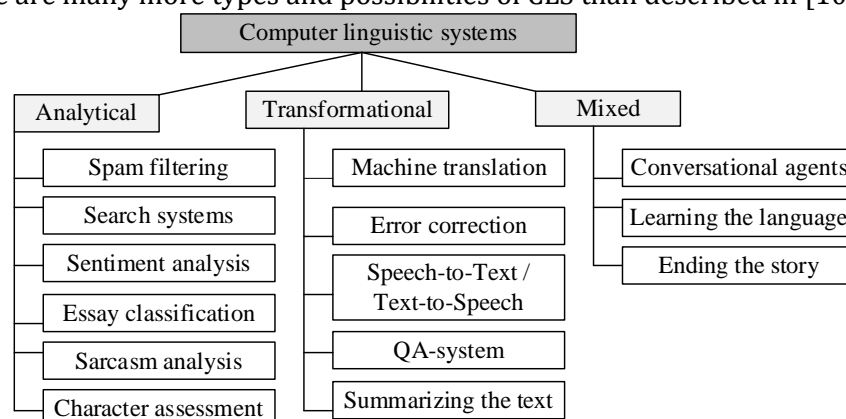


Figure 2: Classification of computer linguistic systems by Grammarly

This list should be supplemented with recommender systems, mass media IS, systems for analysing the psychological state of a person (for example, IBM Watson™ Personality Insights), plagiarism identification systems (copyright/rewrite), systems for determining the author's speech style, speechless access interfaces, sign language recognition systems, etc. Stephen Hawking (one of the most famous people) used a speech computer for communication [16]-[21]. The IBM Watson™ Personality Insights service provides an API for collecting statistical data and corporate information from social networks and other e-communication tools. The service uses linguistic analytics to infer internal personality characteristics of people using e-communication tools such as e-mail, text messages, tweets, and forum posts.

2.3. Basic NLP tasks of computer linguistic systems

The main criterion for market development and the frequency of use of CLS is the motivation for the use of intelligent software, and cloud solutions/applications based on NLP, which improve the service of various customers of all possible areas of human activity and significantly increase the potential audience of modern IT users without the need to possess special skills and knowledge for their use. This was influenced by the range of problems that should be solved by different types/purposes of CLS (Fig. 3) [22]-[27].

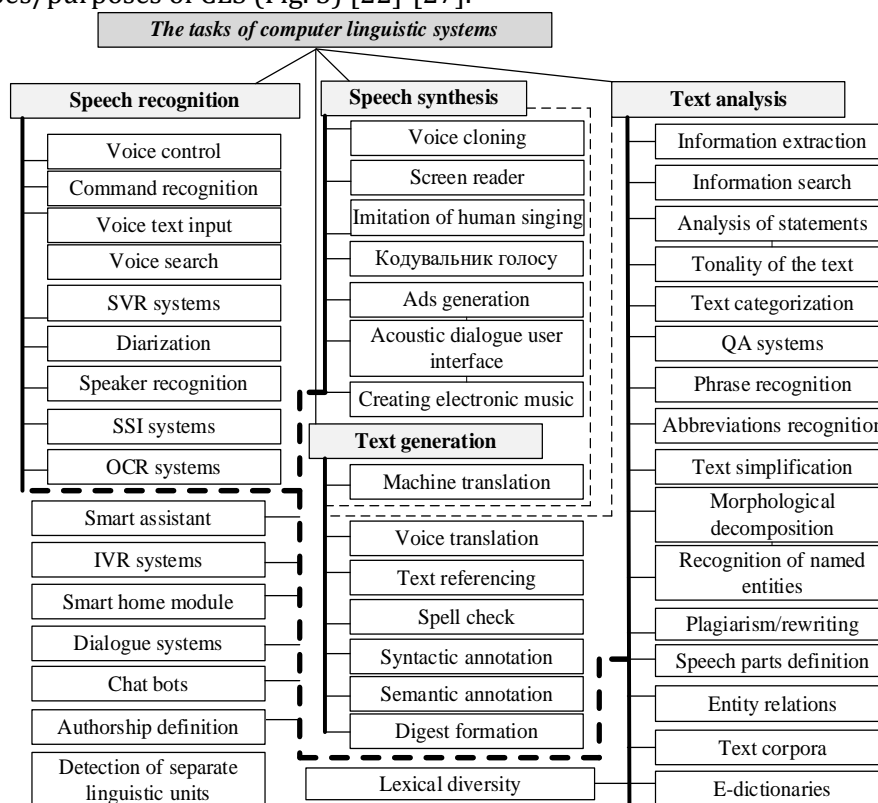


Figure 3: Classification of problems of computer linguistic systems

The main directions of problem-solving for CLS are the analysis and/or generation of texts in natural language, and the recognition and synthesis of natural speech [27]-[33]. Some of the current problems are simultaneously attributed to some directions, for example, dialogue systems rely on such NLP tools as language recognition, content and context selection, identification of intentions, and then building a dialogue based on the above (ideally, by synthesis speech). Thus, a smart assistant should solve the problems of speech recognition, text analysis, text generation and, accordingly, speech synthesis. And machine translation solves the problems of text analysis, speech synthesis and text generation. For QA-systems (question-and-answer), it is sufficient to solve text analysis problems. However, these are only conditional assumptions because each implementation of a specific CLS is usually a closed commercial project, which does

not allow IT specialists to familiarize themselves with the detailed structure of the relevant systems and implemented NLP algorithms.

2.1. Examples and comparative analysis of known modern CLS

Today, with the rapid development of AI, the accelerated growth of large amounts of data and knowledge, and the rapid pace of informatization of society, many CLS (software/Web services) of various purposes have been developed and implemented for solving the appropriate type of NLP tasks according to user needs [34]-[39]. Leading global companies such as Google, Apple, IBM, Microsoft, etc. work in this direction. Along with them, other less well-known companies, including Ukrainian ones, are working on different types of CLS. These CLSs have their advantages and disadvantages. Let us consider and compare only the most popular world and international CLS projects, respectively, from each class of NLP tasks (Table 1). Companies such as NASA, IBM, Apple, Amazon, Microsoft, Google, Yamaha, Grammarly, DARPA, Yahoo, etc. work in the field of computer linguistics.

Most CLS projects are closed, one-off and commercial. Only individual enthusiasts reveal the secrets of their projects and provide users and IT professionals with access to their developments. In addition, most of the developed CLS work with English-language texts, or several European and Asian languages, the list of which does not include the Ukrainian language.

Table 1
Known tools for solving the corresponding NLP problem

NLP task	Speech recognition tool
Speech recognition	
Voice control and command recognition	a component of Microsoft Windows (Vista, v. 7-11), OS/2 Warp 4 and Mac OS X, as well as Voice Access, IBM ViaVoice, Microsoft Voice Command, Yandex SpeechKit, Dragon NaturallySpeaking, Speereo Speech Engine, Lexy, linguistic Voice Pro, Speech (Apple Macintosh), etc.;
Voice input (set) of data	VoiceNavigator, Dragon Naturally Speaking MSpeech (Google Voice API), Voco, Dictate, SpeechPad (Chrome), VoiceNote II (Chrome), TalkTyper (Chrome, 37 languages, free), SpeechTexter (Google Play), Google Docs (Gmail), Voice Notepad (Chrome, 120 languages), Odrey (Ukrainian development), VoiceTypist (voice input in Ukrainian), etc.;
Speech analysis	IBM via voice, Dragon;
Voice IIS	Google, Yandex SpeechKit Mobile SDK, MediaInsight, etc.;
Subvocal recognition (English: Subvocal recognition, SVR) of speech during silence of a person based on the analysis of electromyograms	NASA (Ames Research Laboratory) technology from the Ames Research Center in Mountain View (California), led by Charles Jorgensen, etc.;
Diarization of the speaker to identify the part of the speech and its content to a specific person from the dialogue set	National Institute of Standards and Technology (NIST), etc.;
Speaker recognition (person identification based on voice features for behavioural biometrics	GoVivace Inc., Barclays, Barclays Wealth, CSELT, etc.;
SSI (Silent Access Interface) as an auxiliary tool for creating sound phonation of audio speech or communication in the presence of background noise	AlterEgo (Arnav Kapur, MIT Researcher), SpeakUP (Varun Chandrashekar), NTT DoCoMo, etc.;
OCR (optical character recognition)	ABBYY FineReader, CuneiForm, Brainware, COCR2, ExperVision TypeReader & RTK, Tesseract, FineReaderOnline.ru, FreeOCR, GOOCR, HOOCR, img2txt.com, Microsoft Office OneNote 2007, Microsoft Office Document Imaging, OCRopus, Kirtas Technologies Arabic OCR, NovoDynamics VERUS, NewOCR.com, OnlineOCR.ru, Ocrad, OmniPage, Persian Reader, Readiris, ReadSoft, RelayFax Network Fax Manager, Scantron Cognition, SILVERCODERS OCR Server, SimpleOCR, SmartScore, Tesseract, ViewWise, WeOCR, Zonal, etc..

NLP task	Speech recognition tool
Synthesis of speech	
Voice cloning, voice changing	CereVoice Me, iSpeech, Voice Anonymizer, LyreBird, Resemble AI, Voice changer, Morphvox, SDK packages, Voice Cloning Toolkit for Festival and HTS (Mac, Research Center for Speech Technologies, Junichi Yamagishi from the University of Edinburgh) etc.;
Imitation of a singing person using Vocaloid technology from Yamaha Corporation	SONiKA, LEON, CUL, MAYU, IA, UNI, AVANNA, MEW, DAINA, KAITO, DEX, OLIVER, YANHE, MEIKO, MEIKO V3, LUMi, MATCHA, AZUKI, MAIKA, FUKASE, MIRIAM, LOLA, KAITO V3, CYBER DIVA, Ken, Kaori, Chris, Amy, Haruno Sora, KAITO V5, YANHE V5, CYBER SONGMAN II, CYBER DIVA II, VY1v5, VY2v5, Yuezheng Ling V5, ARSloid, Sachiko, SeeU, Kaai Yuki, Yuzuki Yukari, Luo Tianyi, GUMI Native (Megpoid), Clara, Bruno, Ryuto (Gachapoid), GUMIv3 (Megpoid), VY1 (VocaloWitter), eVY1, VY1 (Mizki), VY1v3 (Mizki), VY1 (i-VOCALOID), VY2 (Yuma), VY2v3 (Yuma), VY1V4, YOHIoid, ZOLA PROJECT, IA ROCKS, CYBER SONGMAN, V flower, v4 flower, Gachapoid V3, Megpoid V4, Rana V4, Hiyama Kiyoteru V4, Kaai Yuki V4, SF-A2 miki V4, Tohoku Zunko V4, Hatsune Miku V4X, Gackpoid V4, Kagamine Rin and Len V4X, Macne Nana V4, Tone Rion V4, Nekomura Iroha V4, Luo Tianyi V4, Megurine Luka V4X, Xin Hua V4, Xin Hua V4 Japanese, Hatsune Miku V4 Chinese, Yuzuki Yukari V4, galaco NEO, Kokone, Ruby, Tohoku Zunko, Tone Rion, Chika, Merli, Lily, Rana, Xin Hua, Tonio, Macne Nana, Aoki Lapis, SF-A2 miki, Megpoid English, anon & kanon, Yuezheng Ling, Hatsune Miku V3 English, Big-AL, Hiyama Kiyoteru, Vocaloid Keyboard, Pocket Miku, Megpoid GUMI, Miku Append, Megurine Luka, Utatane Piko, Nekomura Iroha, Prima, Kamui Gakupo (Gackpoid), Kagamine Rin and Len, Hatsune Miku, Sweet ANN, Kagamine Rin and Len Append, Mo Qingxian, Zhiyu Moke, Mirai Komachi, Kizuna Akari, Yuezheng Longya, Yumemi Nemu, Otomachi Una, Xingchen (Stardust), etc.;
Voice encoder based on vocoder technology and VST-plugins	Cylonix, Darkoder, Lpc-vocoder, Formulator, AC vocoder, Voctopus, Akai DC Vocoder, Fruity Vocoder, Steinberg Vocoder, FL Studio, Steinberg Cubase, Cakewalk Sonar, Buzz Composer, Max MSP, NI Reactor/Generator, etc.;
Screen reader or synthesis of speech from text	VoiceOver (Apple Inc Mac OS X), Арафон, Narrator (MS Windows), Microsoft Agent, GNOME, NVDA, VS Robotics, Window-eyes, JAWS, Festival, AT&T Natural Voices, Gnuspeech, ESpeak, pVoice (Perl), RSS To Speech, Read Words Eng 4, Digalo (Acapela ELAN TTS), Nuance RealSpeak (ScanSoft), Sakrament TTS Engine, Govorilka, Speak Aloud, ToM Reader, CoolReader, Linguattec, Acapela, Oddcast, iSpeech, Google Translate, Зуха, Балаболка, Microsoft Speech Api 4.0, тощо; україномовними є Lesya (KobaVision/KobaSpeech, Code Factory, NextUp, Nuance Vocalizer), Розмовлялка, WaveNet, UkrVox, RHVoive, CyberMova/VymovaPlus/VymovaPro, etc.;
Generating messages/ads	HKUST Xunfei, VS Robotics, etc.;
Acoustic dialogue interface based on Partner-assisted scanning technology (voice output communication aids or Speech-generating devices, SGDs)	Equalizer for Stephen Hawking (Walter Woltosz, CEO of Words Plus), Tony Proudfoot, Roger Ebert, Pete Frates (founder of the ALS Ice Bucket Challenge), etc.;
Creation of electronic music	Adobe Audition, Final Cut Pro (Apple Pro, Mac), MainStage, Logic Pro, Compressor (Apple Qmaster, Apple Qadministrator, Share Monitor), Motion, etc.;
Analysis of text arrays of data	
Information search or information search systems	Google, Yahoo!, Yongzin (China), AltaVista, A9.com (Amazon), LightStorage, Ask.com (Teoma), Alltheweb FAST-Engine, ALLhave, Search engine site ABC Engine, ZipLocal (USA, Canada), Neti (Estonia), Yandex (russia, belarus, Turkey, Kazakhstan), Yahoo Japan (Japan), Walla! (Israel), Seznam (Czech Republic), Sesam (Norway, Sweden), Search.ch (Switzerland), Rediff (India), Rambler (russia), Pipilika (Bangladesh), Naver (Korea), Najdi.si (Slovenia), Miner.hu (Hungary), Maktoob (Arab World), Leit.is (Iceland), Goo (Japan), Fireball (Germany), Egerin (Kurdistan), Daum (Korea), Biglobe (Japan), Accoona (China, USA), Youdao, Yippy, WebCrawler, Swisscows, Startpage.com, Soso, Sogou, Searx, Qwant, Mojeek, MetaCrawler, Lycos, HotBot, Gigablast, Excite, Exalead, Ecosia, DuckDuckGo, Dogpile, Bing, Baidu, Voilà, Nomade, Locace, Francité, Ez2find, Abacho, Wseeker, AOL Search, SAPO (Portugal, Mozambique, Cape Verde, Angola), Google Scholar, Scirus, ArXiv.org, ScienceDirect, PubMed, PDF Search System (PDFSS), LightStorage (media files), GlobalFileSearch (files), Tineye (images), etc.; among Ukrainian search engines, the

NLP task	Speech recognition tool
Expression analysis or content analysis	following should be highlighted: Meta, Shukalka, Bigmir, I.ua, Online.ua, TopPING, UAport, Ukr.net, search.com.uab, etc.; qualitative (Kwalitan, MAXQDA, etc.) and quantitative (TextQuest, Textanz, etc.), WebAnalyst (Megaputer Intelligence), Autonomy Knowledge Server, Text Miner (SAS Institute), InfoStream (Elvista, Ukrainian development), Lithium Community Platform, Meltwater Buzz, etc.;
Development of e-dictionaries	Abby Lingvo, ForceMem, dict, Stardict, GoldenDict, WordNet (semantic English dictionary), ConceptNet, Multitran, Vykislovar, WordNet-Affect, SenticNet, SentiWordNet, etc.;
Text analytics (text mining), information extraction or intellectual text analysis	Intelligent Miner for text (IBM), SAS Text Analytics, WebAnalyst (Megaputer Intelligence), Autonomy Knowledge Server, SemioMap (Entrieva), TextAnalyst (Megaputer Intelligence), Text Miner (SAS Institute), Apache OpenNLP (Java), OpenCalais (Thomson Reuters), Natural Language Toolkit (Python), Galaktika-ZOOM, InfoStream (Елвіста, українська розробка), russian Context Optimizer (RCO), Lithium, Ontos (TAIS Ontos, Ontos SOA, LightOntos for Workgroups, OntosMiner), Paai's text utilities etc.;
Analysis of the tonality of the text	SAS Sentiment Analysis, Lithium Social Media Monitoring, InfoStream (Elvista, Ukrainian development), OpinionEQ, Radian6, OpenAmplify SocialView (Visual Intelligence), Meltwater Buzz, LIQUID CAMPAIGN Opinion Mining, Social Mention, Tweetfeel, Twittratr, etc.;
Identification of key words and persistent phrases (collocations) Text categorization	Feature extraction tool (Intelligent Miner for text, IBM), SemioMap (Entrieva), VICTANA (Ukrainian platform), Oracle Text, InterMedia Text, Galaktika-ZOOM, etc.; Categorization tool (Intelligent Miner for text, IBM), SemioMap (Entrieva), Autonomy Knowledge Server, TextAnalyst (Megaputer Intelligence), RCO, AskNet, etc.;
Text clustering	Clusterisation tool (Intelligent Miner for text, IBM), SemioMap (Entrieva), TextAnalyst (Megaputer Intelligence), Vivisimo Nigma, Quintura Searchcrystal, etc.;
Question-and-answer systems (QA-systems)	ELIZA, Watson (IBM), DrQA (Facebook Research), etc.;
Phrase recognition	WebAnalyst (Megaputer Intelligence), Oracle Text, InterMedia Text, Google Translate, TextGrabber, Translate.Ru, Yandex Translate, Microsoft Translate, Translator Foto - Voice, Text & File Scanner, iA Writer, TextExpander, Odrey (Ukrainian development), Apache OpenNLP, etc.;
Morphological decomposition	Oracle Text, InterMedia Text, iA Writer, TextExpander, Odrey (Ukrainian development), RCO, Apache OpenNLP, etc.;
Recognition of nouns, collocations, idioms, idioms and catchphrases	OpenNLP, SpaCy, GATE, SemioMap (Entrieva), Autonomy Knowledge Server, Oracle Text, InterMedia Text, Galaktika-ZOOM, DBpedia Spotlight, Apache OpenNLP, etc.;
Definition of parts of language words	Oracle Text, InterMedia Text, Apache OpenNLP, etc.;
Language identification	Language identification tool (Intelligent Miner for text, IBM), etc.;
Recognition of abbreviations	OpenNLP, SpaCy, GATE, VICTANA (Ukrainian platform), etc.;
Simplification of the text	WebAnalyst (Megaputer Intelligence), Poetica, Test-the-Text, HamingwayApp, Readability, No stop words, Ilya Bierman's typographic layout (Gagadget, Ukrainian version), Typographer Artemiy Lebedev, LeoBilingua, Forson, etc.;
Identification of plagiarism/rewriting or duplication of text	Unplag/Unichek, Etxt Antiplagiat, Advego Plagiatius, Plag.com.ua, Plagiarisma, Content-watch, StrikePlagiarism.com, TEXT.RU, Edu-Birde, InfoStream (Елвіста, Ukrainian development), etc.;
Definition of relationships between entities	Text Miner (SAS Institute), SemioMap (Entrieva), Autonomy Knowledge Server, Galaktika-ZOOM, InfoStream (Elvista, Ukrainian development), Link Grammar Parser, Mystem, LingSoft, Cíbola/Oleada CLR, StarLing, MCR DLL, SyTech, etc.;
Solution of lexical polynomiality	Oracle Text, InterMedia Text, etc.;
Coreference analysis – definition of a set of terminological nominal entities related to one object, subject, phenomenon or event	TextAnalyst (Megaputer Intelligence), Customer Intelligence Center, etc.;
Statistical analysis of the text	WordStat, netXtract Relevant, URS, FRQDictW, Lemmatizer Multitran, Textarc, Ngram Statistics Package (NSP), Rhymes, WordTabulator, etc.;
Detection of individual linguistic units	Autonomy Knowledge Server, SemioMap (Entrieva), Galaktika-ZOOM, etc.;

NLP task	Speech recognition tool
Marking and labelling of texts for the formation of linguistic corpora of texts	GenCode, TeX, LaTeX, Scribe, GML, SGML, HTML, XML, XHTML, Textual Analysis Computing Tools (TACT), etc.;
Generation of scripts/plots for plays/television programs/films	Final Draft, etc.;
Editor for concentration	FocusWriter, iA Writer, etc.;
Automatic HTML editor	Reformator, etc.;
Creation of concordances	WordSmith Tools, MonoConc, Textual Analysis Computing Tools (TACT), ParaConc, WordSmith Tools Mike Scott, Concordance 2.0.0 R.J.C. Watt, etc.

Generation of text datasets based on speech recognition/synthesis and text analysis

Machine translation	Google Translate, Microsoft Translator, PROMT, SYSTRAN, Yandex.Translate, TIDES, Babylon translator, MT@EC, Trados, OmegaT, Apertium, SDL Trados, STAR Transitfr, Déjà Vu, SDLX, Abby Lingvo, Socrat, Across Language Server, CrowdIn, GlobalSight, gtranslator, MateCat, memoQ, MetaTaxis, Open Language Tools, Phrase, Poedit, Pootle, Babylon, SDL Trados Studio, SmartCAT, UNMIN, Virtaal, Wordfast, Anusaaraka, DeepL, GramTrans, IBM Watson, IdiomaX, Moses, Moses for Mere Mortals, NiuTrans, OpenLogos, etc; Ukrainian development: Trident Software, Pragma, L-Master 98, Language Master; Utility program GoogleTalk, Facebook, MSN Messenger, Skype, etc.;
Identification of search spam (Spamdexing)	Google, Yahoo!, AIRWeb, etc.;
Creating a rewrite	BIPOD, ReWrite Suite, SeoGenerator, korektoronline.pl, Program for rewriting (plati.ru), etc.;
Checking spelling and grammar (spell checker)	Grammarly (developed by Ukrainian programmers), Microsoft Word, myspell, aspell, ispell, Orfo, SPELL (Ralph Gorin, Stanford Artificial Intelligence Laboratory), WordPerfect, WordStar, Firefox, GNU Aspell, Mac OS X, Pidgin, Kmail, Opera, Konqueror, Google, Online Corrector (Ukrainian development), Draft, Google Docs, Orfogrammka, Language Tool (Ukrainian orthography), etc.;
Voice translation	Speereo Voice Translator, etc.;
Referral	Annotation tool (Intelligent Miner for text, IBM), Oracle Text, InterMedia Text, RCO, etc.;
Syntactic annotation	WebAnalyst (Megaputer Intelligence), RCO, LeoBilingua, Forson, etc.;
Semantic annotation	TextAnalyst (Megaputer Intelligence), RCO, Customer Intelligence Center, Ontos, etc.;
Formation of digests	InfoStream (Elvista, Ukrainian development), etc.;
Latent Semantic Analysis (LSA)	patent from Lynn Streeter, Karen Lochbaum, Thomas Landauer, Richard Harshman, George Furnas, Susan Dumais, Scott Deerwester as Latent Semantic Indexing (LSI), etc.;
Spam filtering and email routing	Kaspersky Anti-Spam, Apache SpamAssassin, AntispamSniper (The Bat!), etc.;
Stylometry (classification by style and genre)	Emma, VICTANA (Ukrainian platform), etc.;
Linguometry	netXtract Relevant, WordTabulator, Ngram Statistics Package, Rhymes, Langsoft, VICTANA (Ukrainian platform), etc.;
Glottochronology	VICTANA (Ukrainian platform), etc.;
Evaluation of readability	WebFX Readability Test Tool, Automatic Readability Checker, Readability Calculator, Perry Marshall, StoryToolz, Readability Checker, Word Counter, Joe's Web Tools, progaonline.com, ru.readability.io, copywritely.com, glvrd.ru, Advego, turgenev.ashmanov.com, etc.

Mixed direction of speech recognition/synthesis and text analysis

Smart assistant	Google Assistant, Siri (Apple), Amazon Alexa, Yandex Alisa, Robin (Audioburst), Vani Dialer (Bolo International Limited), Assistant Dusya (UseYoVoice), Marusya (VK.com), Okey Bloknotik (Dmitriy V. Lozenko), MYRI (BlueTo), Cortana (Windows 10), Horynich, AGGREGATE, Tyle (Windows), etc.;
IVR systems (Interactive voice response, interactive voice response, voice menu system)	VoiceNavigator, VoiceKey.IVR, (Customer Engagement Platform), etc.;
Smart home module	Apple Siri, Google Home, Facebook M, Xiao Ai, Amazon Alexa, Microsoft Cortana, Sonos One, Yandex Alice, etc.;
Dialogue system or speech agent, SA	GUS system, CSLU, NLUI, LinguaSys, modules in modern games, Olympus, Nextnova, Quack.com, NADIA, etc.;

NLP task	Speech recognition tool
Creation of chat bots	services SendPulse, Flow XO, ManyChat, Chatfuel, MobileMonkey, ChatbotsBuilder, Botmother, ChatBot.com, etc.;
Determination of authorship of texts	Emma, Lingualyzer, Attributor, SMALT, Anti-plagiarism, Fresh Eye, etc.;
Analysis of the psychological profile of the author	IBM Watson™ Personality Insights, Avtoroved, etc.;
Analysis of scientific literature (determination of novelty and relevance, semantic search, identification of homonyms, etc.)	NaCTeM services (National Center for Text Mining, Manchester and Tokyo Universities), BioText (School of Information, University of California, Berkeley, USA), TAPoR (University of Alberta, Edmont, Canada), etc.

CLS ELIZA (Fig. 4) is one of the first examples of solving the NLP problem of conducting a dialogue between a computer and a user, imitating the answer of Rogersky's psychotherapist [39]-[43].

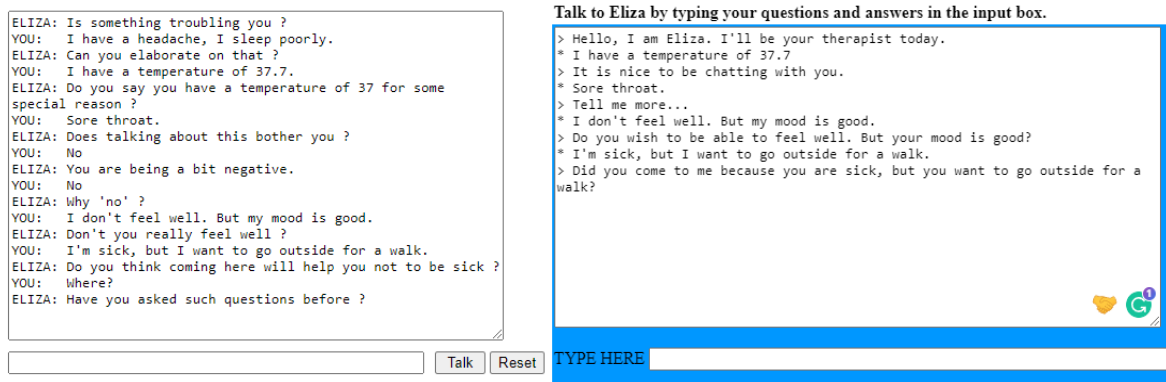


Figure 4: Examples of dialogues in CLS ELIZA

Unfortunately, this dialogue is limited in its structure and intended only for English-speaking participants [39]. ELIZA is a classic CLS that applies a pattern to identify English phrases of the form *You are X* and transforms them into typical questions like *What makes you think I am X?* (where *X* is an arbitrary string of English words from the user). This is an imitation of a dialogue without semantic analysis to understand the content of questions from the user. In [40], the author noted that ELIZA implements one of several dialogue genres where listeners can act as if they know nothing about the world around them. In the beginning, most users of ELIZA came to believe that the system understood them and their problems, even after the explanatory publication [41]-[43]. This is one of the first attempts to implement chatbots, which are now filled with modern social networks, services and e-commerce systems.

Of course, today's conversational agents are much more than entertainment; they can answer questions, book a flight, or find restaurants, the functions they rely on are much more complex to understand than the user's intent [39]. However, the simple, model-based techniques used by ELIZA and other chatbots are critical to solving today's NLP problems. But for chatbots in the Ukrainian language, the usual use of templates makes the process of imitating communication impossible due to the presence of word changes (by case, tense, plural, etc.) depending on the context. Without a simple morphological, lexical and syntactic analysis, constructing an answer or question in Ukrainian in such CLS is not an acceptable result. In addition, regular expressions, text normalization, tokenization, lemmatization, stemming, segmentation, and editorial distance calculation should be used for text templates [44]. Regular expressions are used to identify a sequence of characters to be extracted from previous user queries. For this purpose, word tokenization is used to separate them from the main text (simple identification of word boundaries by the presence of spaces and punctuation marks is an insufficient process for extracting phrases like *проспект Червона Калина* (Chervona Kalina Avenue), *Улан-Уде* (Ulan-Ude), *Алма-Ата* (Alma-Ata), *Південна Корея* (South Korea), *Івано-Франківськ* (Ivano-Frankivsk), *село Залізний Порт* (Zalizniy Port village), *місто Гола Пристань* (Gola Prystan city), *місто Кривий Різ* (Kryvyi Rih city), *Кам'янець-Подільська фортеця* (Kamianets-Podilskiy

fortress), etc. or type abbreviations, і т.п. (etc.), т.д. (etc.), англ. (English), грн. (UAH), and various abbreviations, for example, *CLS*, and *NLP*). Also, today, when tokenizing, you need to take into account various punctuation marks for conveying emotions in the form of emoticons, for example, :), :(, :))) etc. or hash tags such as #lpnu, #friends. The presence of stylistic errors, such as the absence of spaces between words or appropriate punctuation marks, complicates the tokenization process. Text normalization is transforming it into a convenient standard form for the perception of the CLS user, taking into account and matching all inflexions for all words in the sentence. Text normalization also uses segmentation or parsing – breaking the text into separate sentences using punctuation marks as signals. In lemmatization, the same words are identified by analyzing their roots, despite their difference, for example, the words *біжать* (run), *бігли* (ran), *збігли* (ran, etc. are forms of the verb *бігати* (to run). Stemming is a simpler version of lemmatization, where words are shortened to the base by simply dropping suffixes and/or inflexion. For the speed of processing texts in Ukrainian, it is better to use stemming (IP based on key words and stable word combinations), and for the accuracy of the obtained results - lemmatization (identification of plagiarism and rewriting). To compare words with scattered chains of symbols, a metric is used - editorial distance, which determines the degree of similarity of the analysed linguistic units based on the number of necessary edits (insertions, deletions, replacements) to replace one sequence of symbols with another. It is used most often in identifying and correcting errors, determining the level of plagiarism-rewrite, IP, generating a text rewrite, recognizing the language/speech of a specific person and machine translation. The following categories of machine translation systems are distinguished: statistical (Statistical Machine Translation, SMT), based on grammatical rules (Rule-Based Machine Translation, RBMT), and hybrid systems. But in each of them, computational semantics analysis methods are used to convey the specific content of the text when translating it into another language – different ways of modelling the meanings of words, phrases, sentences, and fragments of texts. Computational semantics are divided into distributive, ontological, formal, operational, and traditional. In particular, distributive semantics is used to determine the meaning of a linguistic unit based on the statistics of the combination of words in large text corpora of a certain subject [45]-[54]. In ontological semantics, semantic dependencies of linguistic units of the context are calculated to form a set of knowledge. Formal semantics is used to describe the meanings of expressions through mathematical logic and Boolean algebra. Operational semantics allows you to describe a set of text sentences as a set of commands for controlling some process of generating events or functioning of an executive device. Traditional semantics describes the meaning of linguistic units of the text using special interpretation languages. Each of them has its advantages and disadvantages, especially for syntactic natural languages like Ukrainian.

3. Materials and methods

3.1. Structural diagram of linguistic analysis of textual content

Any text in natural language contains a significant amount of abstract informal unstructured content data. This is a meaningful chain of symbolic (linguistic) units with a set of appropriate properties p_j for solving certain linguistic problems (Fig. 5) [54]-[63], as:

- number of sentences, words, words per sentence, etc.;
- size and placement of paragraphs;
- word length and position of the word in the sentence;
- the number of syllables in a word and the number of word contents;
- ratio of consonants and vowels;
- word depth in the sentence dependency tree;
- N-grams and morphemes: affixes, roots, endings;
- is the word capitalized/hyphenated/compound;
- grammatical categories of different POS, etc.

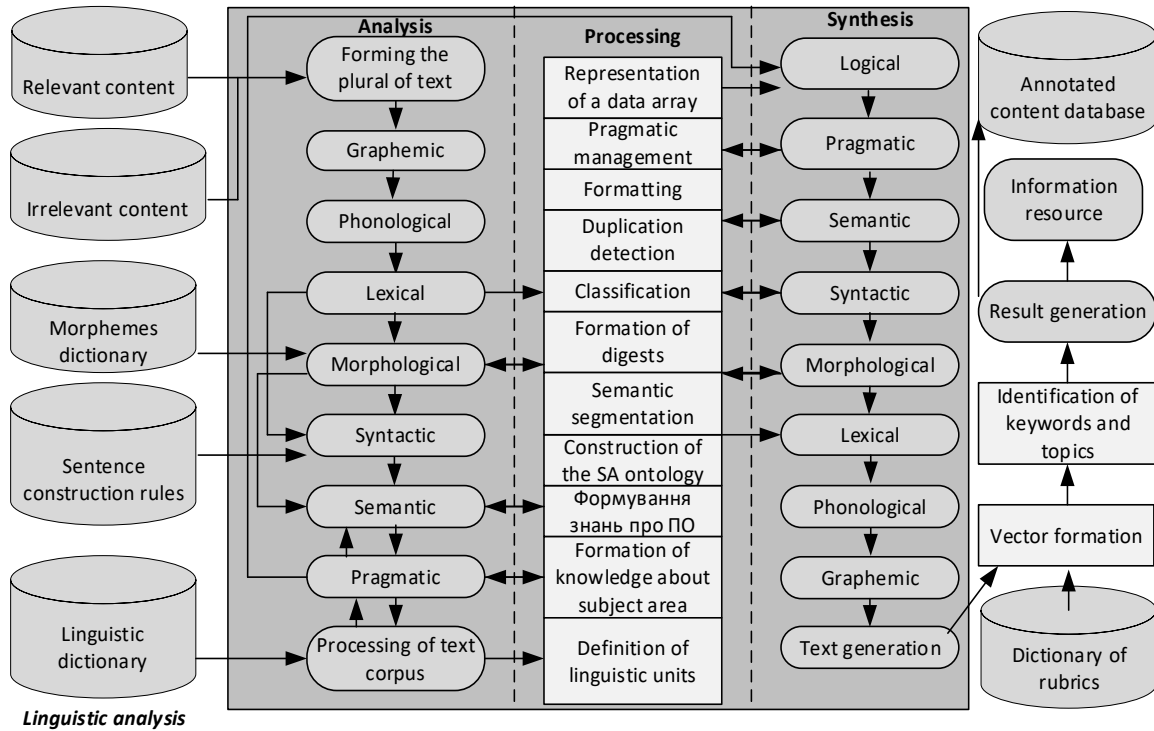


Figure 5: Structural scheme of linguistic analysis of textual content

During linguistic analysis in CLS, different levels of natural language text analysis are used to solve specific problems [56]:

- segmentation – tuples of linear chains of characters delimited by appropriate punctuation marks;
- stemming – sets of linear chains of morphological structures;
- tokenization – a linear sequence of chains of symbols (words, etc.);
- parsing – a network of interconnected structural units in these sentences (grammatical – lexical – phonological categories).

3.2. States and properties of computer linguistic systems

Any CLS state is determined by a tuple of main properties at a specific moment of time or activity of the corresponding NLP process [55]-[59]:

$$s_i = (p_{i1}, p_{i2}, \dots, p_{im}), i = \overline{1, n}, \quad (1)$$

where s_i is the corresponding i -th state at a specific time t_i from the set with power $|S|=n$, p_{ij} is the corresponding ij -th property of the state from the set with power $|P|=m$, which determines the behaviour of CLS:

$$p_j = (r_{ij1}, r_{ij2}, \dots, r_{ijv}), j = \overline{1, m}, \quad (2)$$

where r_{ijk} is the corresponding parameter of the specific property p_{ij} for the state s_i .

For any CLS, the state s_i can be one of the natural language processing processes, for example, the identification of key words and/or stable phrases for the next state s_{i+1} of the system as a rubric of a text array of data. Accordingly, the properties of the state s_i are morphological p_{i1} , lexical p_{i2} and syntactic p_{i3} , in some cases, for the accuracy of the analysis, there may be semantic ones, etc. Then, for the property p_j , a set of parameters will be determined for the corresponding text analysis, depending on the specific NLP task. According to these parameters, the strategy of CLS functioning at the current moment is specified [60]-[66]. For example:

- the parameters of the morphological property p_{i1} are N-grams and morphemes: roots r_{i11} , endings r_{i12} , affixes r_{i13} ; grammatical categories of different POS r_{i14} , word length r_{i15} ,

word location in a sentence r_{i16} , number of syllables in a word r_{i17} , number of word contents r_{i18} , ratio of consonants and vowels r_{i19} , etc.;

- the parameters of the lexical property p_{i2} are the location of the sentence in the test r_{i21} , the location of the word in the sentence r_{i22} , the weight of the word r_{i23} , the weight of the sentence r_{i24} , the base of the word r_{i25} , the inflexion of the word r_{i26} etc.;
- parameters of the syntactic property p_{i3} are the depth of the word in the dependency tree of the sentence r_{i31} , the location of the word in the sentence r_{i32} , the number of contents of the word r_{i33} , the number of words per sentence r_{i34} , the number of words r_{i35} and sentences r_{i36} , whether the word is a capital letter r_{i37} / with a hyphen r_{i38} / compound r_{i39} etc.;
- parameters of the semantic property p_{i4} are the number of word contents r_{i41} , the depth of the word in the tree of sentence dependencies r_{i42} , the size of paragraphs r_{i43} , the placement of paragraphs r_{i44} , paragraph weight r_{i45} etc.

Depending on the property tuple p_j , CLS behaviour is determined, that is, the implementation of a set of rules (activation of actions or events) for the implementation of a specific NLP process to achieve a certain goal depending on the input text data. Accordingly, the event o_l is a change of one property to another $p_{ij} \rightarrow p_{ik}$ or $o_l: p_i \rightarrow p_j$ according to the fulfilment of certain conditions U for the input analysed text X and the intermediate processed text C :

$$p_i = o_l(p_j, U, X, C). \quad (3)$$

Action d_g is the process of activation of an event o_l by another event o_v in CLS:

$$C' = d_g(o_l \circ o_v). \quad (4)$$

The more complex the language (morphology, syntax, etc.), the more difficult it is to automate the processing of relevant texts in natural language. In addition, for languages such as Ukrainian, there are no standardized rules and dictionaries for processing texts in natural language for solving the corresponding NLP tasks. Many scientific linguistic schools and IT specialists are working on the creation of Ukrainian dictionaries and rules for processing Ukrainian texts. But usually, these are linguists and philologists who are not familiar with the features of specific modern tools, such as programming languages, machine learning methods, BigData analysis, etc. There is a colossal gap between the research results of philologists and applied linguists on the one hand, and IT specialists on the other for developing Ukrainian-language tests. In addition, today quite a few NLP tools for the Ukrainian language have been implemented and implemented for public access.

3.3. Classification and features of the main properties of states of the computer linguistic system

Each state s_i CLS for solving a specific NLP problem uses several or all levels of NLP processes to form a tuple of main properties $s_i = (p_{i1}, p_{i2}, \dots, p_{im})$, $i = \overline{1, n}$ [67]-[72]. The content of the speech of any person, regardless of the specific presentation of information (written or audio), is transmitted by each of the six properties p_{ij} of the NLP process or the level of analysis of human language, regardless of origin (Fig. 6):

$$S_{LA} = d(p_I, p_{II}, p_{III}, p_{IV}, p_V, p_{VI}). \quad (5)$$

I. Phonological level	Organization and interpretation of speech sounds
II. Morphological level	Identifying and analyzing word structure and form
III. Lexical level	Division into chapters, paragraphs, sentences, words
IV. Syntactic level	Words analysis as grammatical structure of sentence
V. Semantic level	Determining the sentence meaning in text context
VI. Pragmatic level	Interpretation of sentences in appropriate contexts

Figure 6: Classification of the main subprocesses of natural language processing

The main NLP tasks are closely related. Therefore, some of the processes in CLS for solving various NLP problems are similar or even partly the same. For example, for NLP tasks of machine translation, correcting grammatical errors, identifying keywords, text classification, etc., it is necessary to apply morphological and syntactic analyses of the text. The process of categorizing the text necessarily includes the process of defining keywords. The processes of abstracting and semantic annotation include not only morphological and semantic analysis but also semantic analysis and definition of keywords. Any text analysis should include the lexical level. In addition, each level of text analysis for solving a specific NLP problem can consist of different sequences of steps and their number. NLP methods are used for relevant analyses within the framework of the solution of a specific NLP problem. For convenience, the main subprocesses of NLP are divided into linguistic categories, which are solved by certain methods (Fig. 7). Rules for processing textual content are generated according to these methods. For more effective NLP content, it is necessary and sufficient that CLS implements the maximum possible number of modules of relevant speech levels for solving a specific NLP problem. Each of the relevant levels of text content analysis has its own set of methods for achieving effective specific results depending on the language of the text.

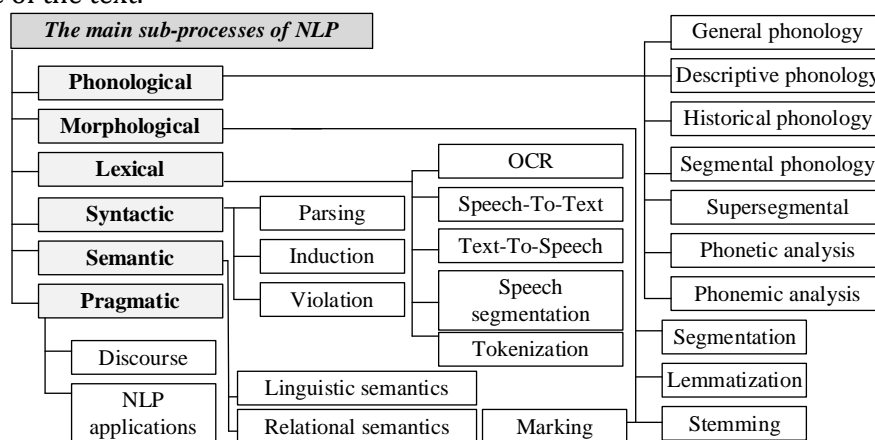


Figure 7: Classification of the main methods of natural language processing

For example, when determining a set of keywords in an English text, parsing, stemming, and various statistical methods are used to analyse the frequency of use of noun group words and their distribution in the text. Accordingly, for Ukrainian-language texts, when defining keywords, the simple stemming algorithm must be replaced with a modified stemming algorithm due to the presence of a large number of inflexions in the analysed text to identify the nominal group. In addition, it is not necessary to compare words, but the bases of noun group words, since in the Ukrainian language keywords are often determined not only by the sequence of words, but their mutual permutation in different cases is possible (for example, *пошук інформації* (information search) – *основи пошук інформац*, *інформаційний пошук* (information search) – *основи інформац пошук*, *пошуку інформації* (information search) – *основи пошук інформац*, etc., that is, cutting off not only inflexions but also suffixes to bring to the base of the word).

3.4. Classical approaches and trends in natural language processing

Induction, deduction, the method of hypotheses, analysis and synthesis, observation, idealization, modelling, and formalization are used to study natural language. In addition, specialized approaches are used to study the phenomena and regularities of a specific natural language as an object of computer linguistics (Fig. 8). These approaches make it possible to define a set of procedures and algorithms for the analysis of speech phenomena to solve a specific problem and, accordingly, to check the obtained results during experimental testing. Usually, for a specific NLP task, a hybrid approach is used as a combination of several different approaches (Fig. 8). For example, the methods of statistical analysis, probabilistic modelling and ML, along with the linguistic approach, are used to determine the authorship of a text, the stylistics of an

individual author, in deciphering, shorthand, language didactics, abstracting, removing polysemy and IP. Statistical methods are used in content analysis to identify the state of social consciousness or emotional colouring to promote relevant political and/or commercial advertising in social networks.

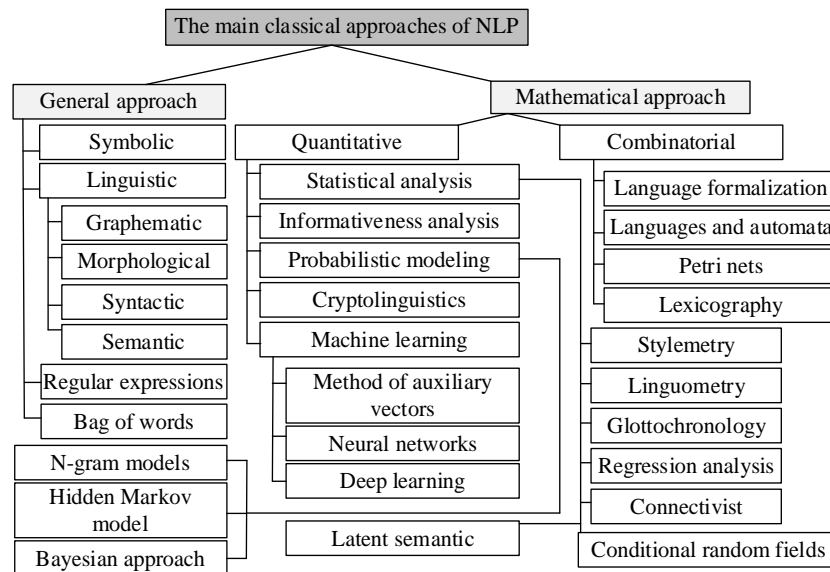


Figure 8: Classification of the main approaches to natural language processing

In linguistic monitoring, in addition to the listed set of methods, regular expressions and a bag of words are used to study the functioning of language in a specific scientific, political or mass media discourse. The purpose of monitoring is also the identification of foreign language borrowings, plagiarism/rewriting, grammatical/stylistic errors, vocabulary of emotions/feelings, thematic/spatial/temporal vocabulary, etc.

3.5. General classification of research directions for NLP problems

Appropriate approaches are used to solve specific NLP problems in typical CLS systems in the appropriate areas of research (Fig. 9). But when solving each NLP problem, when applying specific approaches, depending on the research language, different sets of tools are used to successfully and effectively achieve the set goal. For example, the analysis and identification of psychological effects laid down by the author of textual content depend on the availability of a personalized dictionary of the author and a sentiment dictionary of this region (not all words have the same emotional colours and in different languages and different regions, even different people of specific people - a simple translation will not help to get a real description of a person's psychological state). For example, according to the BigFive model, 5 indicators of a person's psychological state are determined based on his comments on social networks for a certain period, in particular, levels of Extraversion, Introversion or Ambiversion, Benevolence, Agreeableness, openness to experience, Openness, Neuroticism, and Conscientiousness. To analyse the level of Extraversion, lexical measures are studied in the form of an analysis of the set of marker words used in the texts, which respectively reflect the features of a specific type. One set of markers is classified as active, sociable, talkative, sociable, sociable, and another set as reserved, quiet, passive, thoughtful, etc. Markers can be not only adjectives and nominal groups, but also verb groups in a certain tense as a description of actions in time (active or, accordingly, passive). At this stage, complexity arises in syntactic and semantic analysis, depending on the language of the author of the text. In the English-language text, especially in spoken dialogues, there is a clear order of word groups (noun, verb), compared to Ukrainian texts. In addition, the average sentence length is significantly shorter in the English-language text. Therefore, it is easier for them to build a syntactic tree of dependencies to analyse the meaningfulness of markers, and not just their presence (as in the well-known quote from a fairy tale - where in the phrase

someone *казнить нельзя помиловать* (to be punished cannot be pardoned); the presence of a punctuation mark in the appropriate place will determine the level of agreeableness of the author of the catch phrase).

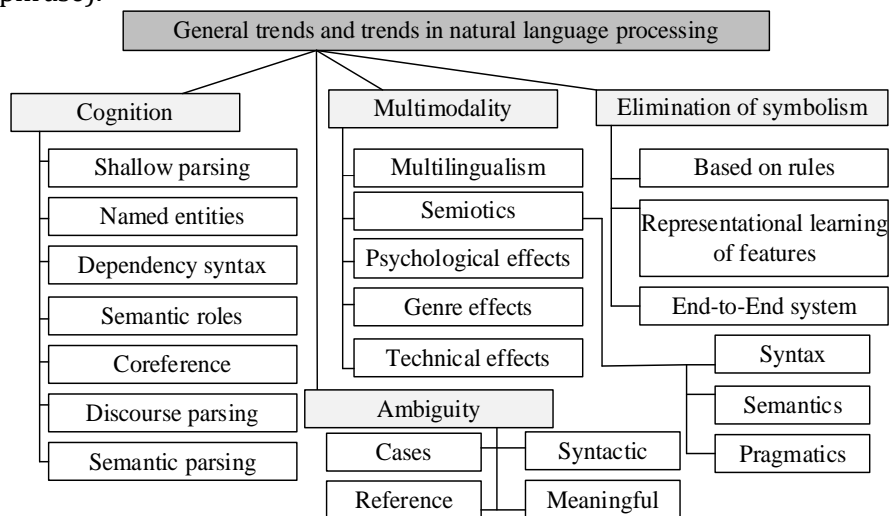


Figure 9: General classification of research directions for NLP problems

3.6. Additional methods of linguistic research for NLP tasks

Additional methods are used for higher-level NLP applications (Fig. 10). The cognitive-onomasiological analysis identifies motivators and the motivational base of phraseological units for the interpretation and modelling of the SA knowledge structure of textual information and the semantic dependence between the motivator and the phraseological unit of the Ukrainian language. The descriptive method is used as part of PHA (phonological analysis), LA (lexical analysis), MA (morphological analysis) and SYA (syntactical analysis) as inventory l_1 , segmentation l_2 , taxonomy l_3 and interpretation l_4 :

$$L = l_4(l_{41}, l_{42}) \circ l_3 \circ l_2 \circ l_1. \quad (6)$$

The internal interpretation of l_{41} groups linguistic units according to multiple criteria. The external interpretation of l_{42} illustrates the connections of a linguistic unit with a meaningful phenomenon, objects, subjects and simulated events of specific text streams of the content of the corresponding language.

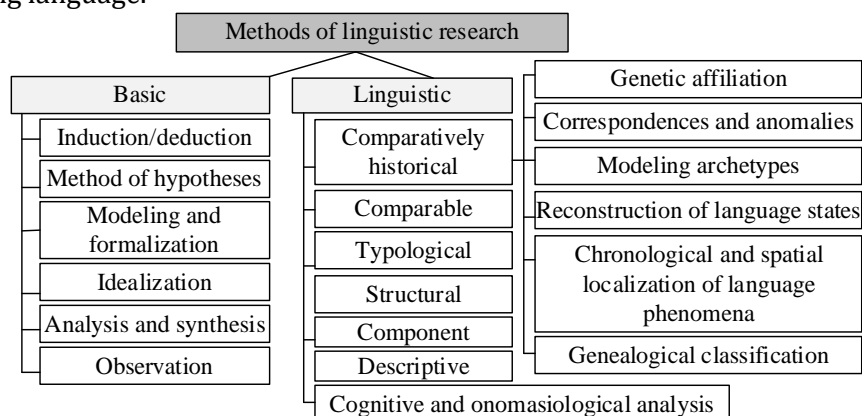


Figure 10: Methods of linguistic research for NLP tasks

The comparative-historical method is used to analyse the kinship of languages based on external (attraction of data) and internal (correlation of phenomena) reconstruction, linguistic statistics, and linguistic geography. The comparative method is used to identify the specific and common characteristics of the analyzed speech texts in the grammatical dictionary and sound

systems based on the comparison to form a criterion as a benchmark for comparing the internal form, onomasiological structure, and word-forming types at the word-forming level, the component composition of the values of comparative equivalents at the lexical level for machine translation systems, dialogue systems and chatbots.

The typological method is used to identify and group the main linguistic features and regularities of speech (clustering) based on differences and similarities of linguistic characteristics. For research, reference language is used, for example, syntactic, morphological and phonetic models, semantic field, grammatical rules, linguistic category, specific language, artificial language, etc.

The structural method is used to study the structure of speech in the methods of transformational, direct components, distributive and component analysis. Syntagmatic, paradigmatic, and epidigmatic relations between sentences, grammes, lexemes, morphemes, and phonemes are analysed. Distributional analysis identifies features and functional characteristics of linguistic units taking into account the environment (distribution). Analysis by immediate components is based on the alternate division of a linguistic unit (*sentence* → *phrase* → *word*) into components until the moment when we get indivisible parts. Transformational analysis identifies semantic and syntactic differences and similarities between linguistic units through features in sets of their transformations when studying lexical semantics, word formation, morphology and syntax. Component analysis is used to determine the lexical meaning of a word as seven (reference-organized set of elementary meaningful lexical units) for the formation of explanatory dictionaries.

Mathematical (statistics and regularities), psycholinguistic (associative experiment, big five), sociolinguistic (analysis through questionnaires), etc. Quite interesting results are given by the methods of psycholinguistic analysis as an associative experiment (free, directed or chain) through the semantic differential. The latter is a qualitative and quantitative indexing of the meaning of the word through two-pole scales with gradation by a pair of antonymic adjectives.

3.7. Research methods of cognitive linguistics

Cognitive linguistics combines knowledge and research in psychology and linguistics with IT tools and AI methods. CL approaches are generative grammar (Generative grammar, author Avram Noam Chomsky), cognitive linguistics (Cognitive Linguistics or linguistics framework, author George Philip Lakoff) [73] and integrative cognitive linguistics (Integrative cognitive linguistics or cognitive semantics). According to George Philip Lakoff, CL is divided into the theory of conceptual metaphor (Conceptual metaphor theory or the analysis of metaphors) and cognitive and construction grammar (Cognitive and construction grammar or the analysis of constructions as in form-meaning with comparison with memes as units of language/speech evolution). George Lakoff proposes a methodology for building NLP algorithms from the perspective of cognitive science, together with the findings of cognitive linguistics, with 2 aspects:

1. Apply the theory of conceptual metaphor to understand one content of a linguistic unit (word, phrase, sentence or text fragment) based on another to identify the author's intention.
2. Assign relative measures of meaning to the analyzed linguistic unit based on the information presented before and after the piece of text being analyzed, for example, using a probabilistic context-free grammar (PCFG). The mathematical equation for such algorithms is given in US patent 9269353 [74]:

$$w(t_N) = p(t_N) \times \frac{1}{2d} \left(\sum_{i=-d}^d (p(t_{N-1}) \times f(t_N, t_{N-1}))_i \right), \quad (7)$$

where w is a relative measure of value; t is a token, any block of text, sentence, phrase or word; N is the number of analysed tokens; p is a probabilistic measure of value based on corpora; d is the location of the marker along the sequence of $N - 1$ tokens; f is a language-specific probability function.

According to the classification of L.A. Kovbasiuk CL is divided into the following directions [73]-[74]:

1. Cognitive poetics – the study of cognitive processes based on which a text array of data is produced, perceived and interpreted;
2. Frame semantics examines cognitive models and mental spaces (frames);
3. Conceptual metaphor and conceptual metonymy (analysis, identification and interpretation of content based on another);
4. The theory of semantic prototypes (category structuring and identification of components based on a given prototype, for example, the prototype of *собак* (dogs) is a *вівчарка* (shepherd) or *маламут* (malamute), of *котів* (cats) is a *Шотландський висловухий* (Scottish short-eared), of *птаців* (birds) is *орел* (an eagle), etc.).

Approaches to the development of cognitive models are used in cognitive, functional and constructive grammar, computational psycholinguistics and cognitive neuroscience (for example, ACT-R or Adaptive Control of Thought-Rational - adaptive control of thought-rationality, authors Christian Lebiere and John Robert Anderson from Carnegie Mellon University). Research directions of cognitive NLP are part of the cognitive AI approach, including based on neural models for multimodal NLP.

3.8. Classification of the main ML methods for NLP processes

The clustering and classification of large text arrays of data is usually carried out based on ML methods (Machine Learning) and big data analysis [67]-[72]. To build such methods, the tools of graph theory, probability theory, optimization methods, mathematical analysis, numerical methods, mathematical statistics, and various techniques of working with data in e-form are used. CLS based on machine learning consists of the main parts as NLP, clustering and classification (Fig. 11).

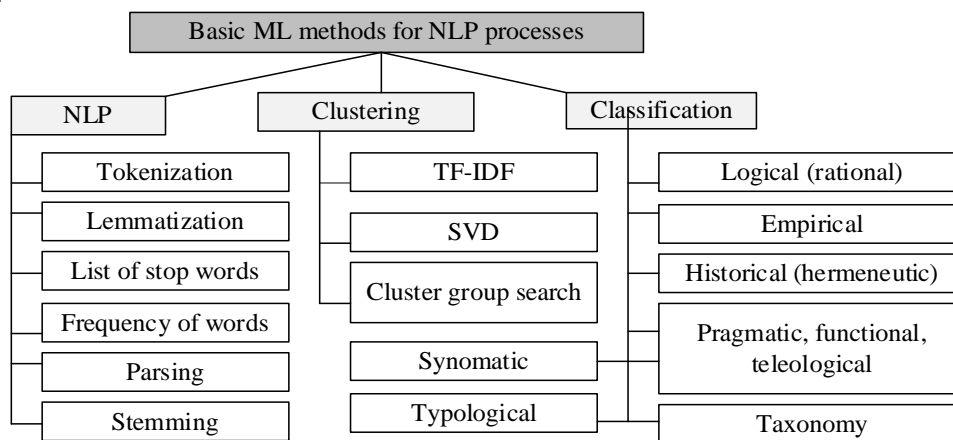


Figure 11: Classification of the main ML methods for NLP processes

Researching texts is one of the most difficult tasks for programmers due to the ambiguity of the meaning of words. Some companies, such as Alchemy and Thomson Reuters, have developed NLP services and ML algorithms for identifying text content. The company Aylien has offered its API toolkit for text analysis for the possibility of creating various NLP services.

The API makes it possible to quickly identify headings and main text in a document, highlight the content and main concepts, and create an abstract or abstract. The data extracted from the text is stored in JSON format, and Mashape is used to provide access to it. Unfortunately, the tool works only with English and German languages.

The main NLP tasks are the development of algorithms for extracting and analysing features of linguistic units of measurement from language and applying them to solve a wider range of CL tasks. Such signs, in particular, are:

- number of sentences, words, words in sentences, etc.;
- size and location of paragraphs;

- the position of the word in the sentence and the length of the word;
- ratio of vowels and consonants;
- the number of syllables in the word and the meanings of the word;
- word depth in the sentence dependency tree;
- composition of morphemes: affixes, roots, endings;
- N-grams and grammatical categories of various POS;
- word with a capital letter / hyphenated / compound.

3.9. The main problems in processing Ukrainian-language texts

The main problems for the development of CLS for studying the Ukrainian language are splitting of linguistic units, marking of parts of speech (POS tagging), parsing and pragmatics, that is, how the context affects the content. Pragmatics studies such features as implicature (ambiguities of statements, hints, guesses), speech acts, relevance and conversation.

Accordingly, the Workflow for NLP for a typical task is:

1. Research available data and NLP algorithms;
2. Prepare test set and baseline and define metrics;
3. Develop an NLP algorithm: feature design; NLP-pipeline (debugging the flow/source of information as a pipeline); NLP resources; choose a rule-based/statistical/ML approach;
4. Implement and test solutions;
5. Monitor execution.

An interesting NLP task is to identify or generate viral news headlines in social networks or online newspapers. There are several features that a potential viral headline should possess:

- uniqueness of the name - lack of analogue;
- proximity – the presence of a reference to the country/city/institution/region of the news source;
- superlativeness – with an indication of the scale/scope or strength/quality of the impact on the phenomenon/subject/object/environment;
- emotion (sentiment) – use of emotional colouring of language;
- surprise – use of unusual phrases/phrases;
- prominence – the presence of a reference to prominent persons (people, locations, titles) or events/actions of these persons.

Semantic text analysis is one of the key NLP problems as a theory of CLS creation. The results of semantic analysis are used to solve problems in such areas as, for example: automatic translation systems (Google translation); IIS (Google is completely based on semantic analysis); philology (analysis of author's texts); trade (analysis of demand for certain products based on comments on this product); political science (prediction of election results); psychiatry (for diagnosing patients), etc. Visualization of the results of semantic analysis is an important stage of its implementation, as it can ensure fast and effective decision-making based on the results of the analysis. An analysis of publications on the Latent Semantic Analysis (LSA) network shows that the visualization of the analysis results is carried out in the form of a two-coordinate semantic spatial graph with plotted words and coordinate documents. Such a visualization does not allow us to identify groups of related documents and to assess the level of their semantic connection by words in the text. For groups of words and documents without visualization, only cluster labels and centroid coordinates were determined.

4. Experiments, results and discussions

4.1. Features of intellectual analysis of content flow

In works [77]-[81] attention is focused on the relevance and perspective of the integration of information flows based on pragmatic text mining methods for solving several TDZ tasks, in particular, abstracting, analysis of information portraits, content analysis of texts, formation of

digests, IIS, etc. This is quite an informative work, but it does not reveal the specifics of processing the texts described in the different languages. So, for the phrase *content analysis*, you will hardly find its other variant *analysis of content* in the English test. In the Ukrainian text, for the keyword *контент-аналіз* (content analysis), there are often used equivalents such as *контентний аналіз* and *аналіз контенту* and their analogues *змістовний аналіз*, *аналіз змісту*. This complicates the process of semantic analysis for NLP tasks of extracting information from textual content, which gives inaccurate resulting data. The process of identification/marketing of keywords/terms, IIS by keywords, and the integration of information flows is significantly complicated, as textual content in the Ukrainian language can take on different forms due to declension with changing inflexions depending on the gender/plural of the noun and adjective, the presence of suffixes, alternation of letters when changing words, etc.

In [82]-[86], for the effectiveness of IIS, it is better to use an ontological approach for English-language content. It is quite effective in extracting knowledge from Ukrainian texts only if detailed and correct morphological, grapheme, lexical and semantic analyses are carried out beforehand. It is possible to build an ontology only with the correct definition and appropriate marking of all connections between all entities with their further preservation in forms (for verbs in the indefinite form - infinitive, for nouns in the nominative form of the singular and adjectives in the form of the nominative case of the masculine gender). That is, for the sentence – *комп'ютерна лінгвістична система розв'язує конкурентну задачу опрацювання природної мови* (the computer linguistic system solves the competitive task of processing natural language), the corresponding analogue will be in the grammatical tree of dependencies in the form of leaves according to the syntactic analysis *комп'ютерний лінгвістичний система розв'язувати конкурентний задача опрацювання природний мова*. Without analyzing this tree, it is impossible to identify word dependencies and their subordination in sentences to build a corresponding ontology automatically based on pressing. In [82]-[86] attention is focused on the relevance and perspective of the analysis of changes in text content streams based on linguistic analysis, including for IIS information in the form of text content. The processing of text information is presented as a set of operators for the formation, management, and maintenance of a set of text content C . Similarly, to previous works, all methods are given for processing content without reference to the specifics of a specific language. In most publications on the features of IIS information, it is recommended to consider the set of analysed content C as a set of subsets of relevant C_{rt} and irrelevant C_{rf} content, or found C_{st} and not found C_{sf} content, or useful content C_{ut} for the end user and useless C_{uf} , or frequently visited content by users C_{vt} and rarely visited C_{vf} or the time of viewing the content is greater than a certain value C_{pt} or less than C_{pf} :

$$C = C_{rt} \cup C_{rf} = C_{st} \cup C_{sf} = C_{ut} \cup C_{uf} = C_{vt} \cup C_{vf} = C_{pt} \cup C_{pf}. \quad (8)$$

The IIS result of the text content is evaluated according to the relevant criteria as the degree of relevance k_1 , relevance k_2 , popularity k_3 , credibility k_4 , uniqueness k_5 , etc. Table 2 presents formulas for determining IIS performance criteria. Each of the listed criteria has its rating scale for forming the rating of the IIS result [87]. The exact calculation of each of the criteria in a specific IIS result does not improve its quality. Improving the value of one of the criteria will lead to a deterioration of the other. Finding a balance of IIS criteria values is a laborious process and does not yield any positive results. But knowing which of the indicators is better to use for a specific purpose of IIS will make it much easier to get the expected result.

Table 2
Criteria for forming the result of IIS text content based on [87] and authors research

k_i	Name	Content	Formula
k_1	Relevance	Correspondence of the number of keywords n of the found content to the number of keywords N of the IIS request and M of the found content, or the ratio of useful to the user to the total	$\frac{n}{N}$, $\frac{n}{M}$ or $\frac{n^2}{MN}$ for specific content, but for everything found then $\frac{ C_{ut} \cap C_{st} }{ C_{uf} \cap C_{sf} + C_{ut} \cap C_{st} }$
k_2	Popularity	The ratio of frequently visited and overtime content to all relevant content found	$\frac{ C_{vt} \cap C_{pt} }{ C_{rt} \cap C_{st} + C_{vt} \cap C_{pt} }$

k_i	Name	Content	Formula
k_3	Topicality	The ratio of frequently visited user-friendly content to all relevant content found	$\frac{ C_{vt} \cap C_{ut} }{ C_{vf} \cap C_{st} + C_{vt} \cap C_{ut} }$
k_4	Certainty	Correspondence of the content to the real meaning and reliability of the source p_s	$\frac{k_1}{p_s}, p_s = [0; 1]$
k_5	Uniqueness	The indicator of the originality of the author's content data with the found ones	$\frac{k_{1i} \cdot C_{st} }{\sum_{i=1}^{ C_{st} } k_{1i}}$
k_6	Authenticity	Indicator of compliance with the source of origin and content authorship	$\frac{k_{1i} \cdot C_{st} }{\sum_{i=1}^{ C_{st} } k_{1i}} \cdot p_s, p_s = [0; 1]$
k_7	Profitability	The ratio of the number of returns $ C_{rp} $ to the total number of views of the content	$\frac{ C_{rp} \cap C_{st} }{ C_{vf} \cap C_{st} + C_{vt} \cap C_{st} }$
k_8	Completeness	The ratio of found relevant content to all possible relevant content	$\frac{ C_{rt} \cap C_{st} }{ C_{rt} \cap C_{st} + C_{rt} \cap C_{sf} } = 1 - k_{13}$
k_9	Precision	The ratio of found relevant content to all found content	$\frac{ C_{rt} \cap C_{st} }{ C_{rt} \cap C_{st} + C_{rf} \cap C_{st} } = 1 - k_{10}$
k_{10}	Noise	The ratio of found irrelevant content to all found content	$\frac{ C_{rf} \cap C_{st} }{ C_{rt} \cap C_{st} + C_{rf} \cap C_{st} } = 1 - k_9$
k_{11}	Precipitate	The ratio of found irrelevant content to all irrelevant content	$\frac{ C_{rf} \cap C_{st} }{ C_{rf} \cap C_{sf} + C_{rf} \cap C_{st} } = 1 - k_{12}$
k_{12}	Specificity or selectivity	The ratio of irrelevant content not found to all irrelevant content	$\frac{ C_{rf} \cap C_{sf} }{ C_{rf} \cap C_{sf} + C_{rf} \cap C_{st} } = 1 - k_{11}$
k_{13}	Remainder, loss or silence	The ratio of no relevant content found to all possible relevant content	$\frac{ C_{rt} \cap C_{sf} }{ C_{rt} \cap C_{st} + C_{rt} \cap C_{sf} } = 1 - k_8$
k_{14}	Uncertainty	The ratio of not found relevant content to all possible irrelevant content	$\frac{ C_{rt} \cap C_{sf} }{ C_{rf} \cap C_{sf} + C_{rt} \cap C_{sf} } = 1 - k_{15}$
k_{15}	Ambiguity	The ratio of irrelevant content not found to all possible irrelevant content	$\frac{ C_{rf} \cap C_{sf} }{ C_{rf} \cap C_{sf} + C_{rt} \cap C_{sf} } = 1 - k_{14}$
k_{16}	Pertinence	The ratio of the amount of useful content according to the user's needs to the total amount of content found	$\frac{ C_{ut} }{ C_{st} }$
k_{16}	Conformity	Relevancy of useful content to found content	$\frac{\sum_{i=1}^{ C_{st} } k_{1i}}{ C_{st} \cdot C_{ut} }$
k_{17}	Satisfaction	Performance indicator for integrated assessment of IIS	$k_8 + k_9$
k_{18}	Functionality	IIS quality assessment performance indicator	$k_8 \cdot k_9$
k_{19}	Specification	The ratio of the accuracy coefficient to the probability of relevance of random content p_{rc} in the array	k_9/p_{rc}
k_{20}	Convertibility	Failure ratio $ C_{cv} $ to the total number of views of the content	$\frac{ C_{cv} \cap C_{st} }{ C_{vf} \cap C_{st} + C_{vt} \cap C_{st} }$
k_{21}	Novelty	Performance indicator for evaluating the quality of found relevant content	$k_3 \cdot k_5$

If the result is to satisfy the end user in a particular PPI, this is one set of criteria, and even in it, there are more important criteria and less important ones (the relevance indicator can prevail over the uniqueness indicator, and to calculate the relevance, accuracy and completeness criteria are used, taking into account the reading time and frequency visits to the candy store by previous visitors). If it is necessary to identify a set of Websites where some grey/black SEO methods are used, then another set of criteria is used, in particular, noise and sediment. One hundred per cent effectiveness of IIS is impossible due to the subjectivity of the author's content, the presence of noise due to the use of grey/black SEO technologies for website promotion, and the incorrectness of creating content search patterns (CSP) due to the complexity of linguistic processing of languages, in particular, Ukrainian.

4.2. Technologies of intellectual analysis of text flow

One of the widespread IT analyses of the flow of textual content is the integration of data from different sources (Fig. 12). Data from reliable sources is usually integrated by tag analysis.

However, the complex process of integration is based on extracting information or data from different sources of content with the growth of NLP methods. Qualitative generation of new textual content from a set of different in nature, but similar in content, data from different sources is one of the most relevant and promising NLP tasks today, for example, for successful e-business management. The stages of generation and application of a set of text content determine the methodology of collection, filtering, indexing, formatting, structuring of information from relevant sources, and further storage, processing, support, formation, management, etc., that is, the main stages of intellectual analysis of the flow of text content (Fig. 13). The process of intellectual analysis of text flow consists of:

1. content integration based on text recognition and analysis;
2. content management based on text analysis and processing;
3. content support based on analysis and synthesis of information.

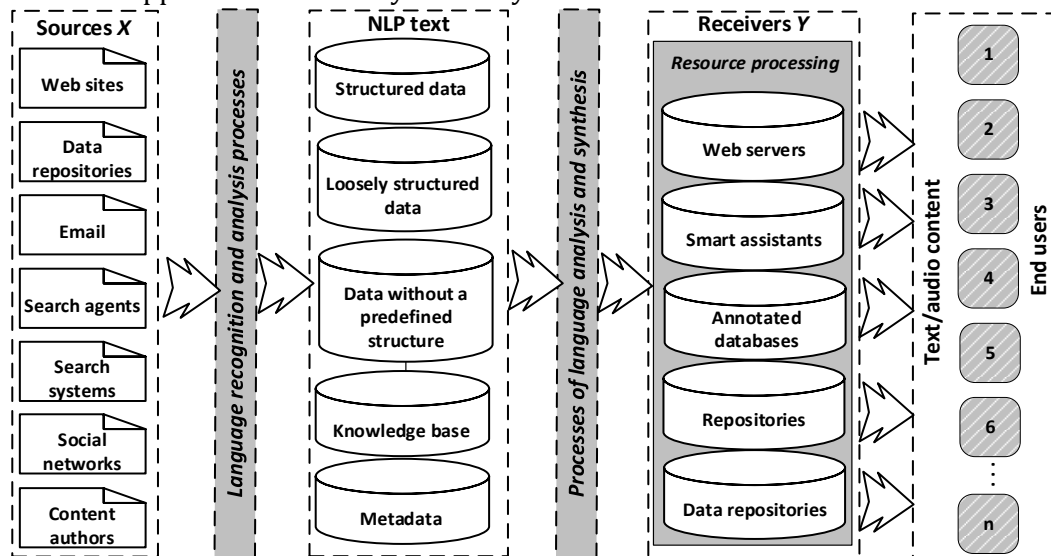


Figure 12: The process of integrating test data from X sources into Y receiver

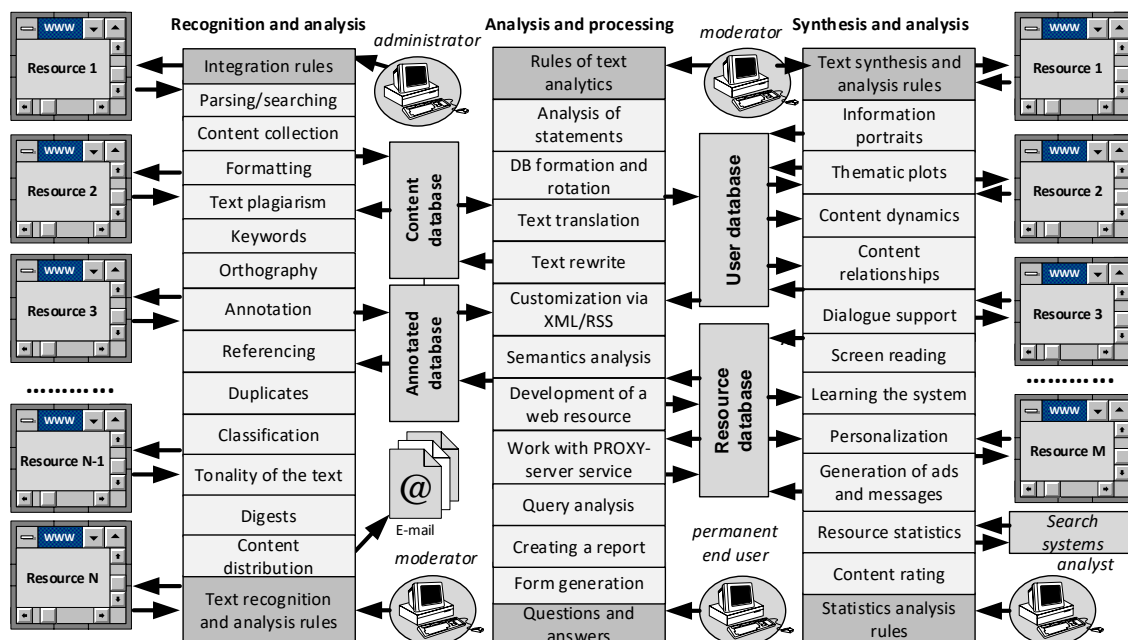


Figure 13: General scheme of intellectual analysis of the text stream

4.

The process of content integration ensures the formation of content based on the methods of content monitoring, content analysis, information extraction and speech recognition from

various sources according to the information needs of the permanent/potential audience (Fig. 14), in particular:

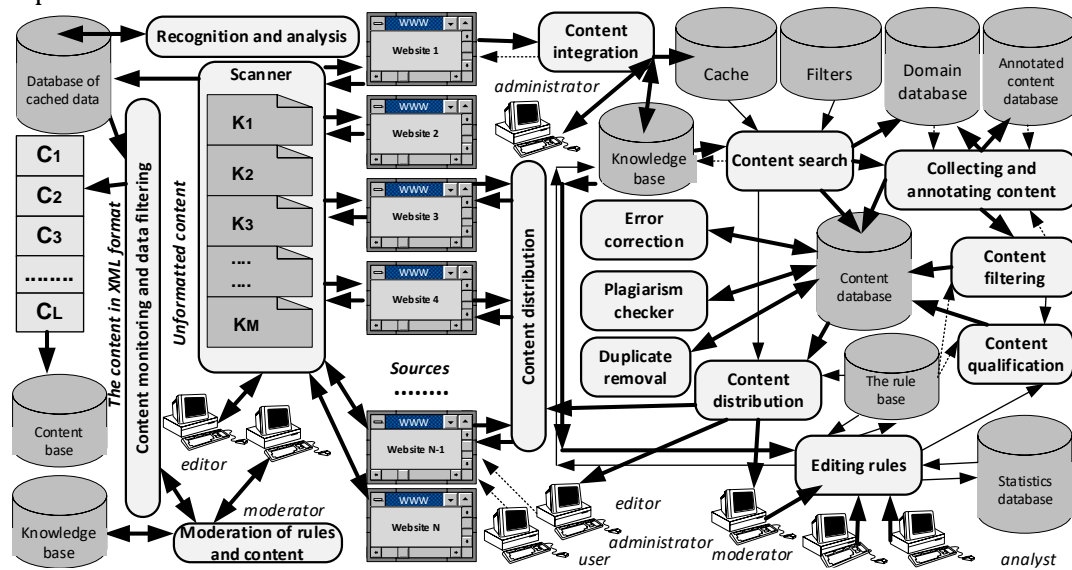


Figure 14: Scheme of integration of text content from different sources

administrator → integration rules → knowledge base → formation of a set of IIS parameters → content database → IIS by parameters → database of cached data → source parsing → information extraction → annotated content database → content formation → content database → content distribution → **moderator**

moderator → rules of text recognition and analysis → knowledge base → content formation from integrated data → content database → content systematization → content database → content distribution → **editor** → content publication → **Website/Web page**

integrated data → content collection → cached database → content formatting → content database → plagiarism check and removal → content database → duplicate removal → error correction → content database → keyword identification → annotated content database → content annotation → annotated database content data → abstracting → content database → rubrication → content database → sentiment analysis → content database → digest formation → content database → content distribution → **potential/permanent audience**

The content management process is described in the following terms:

User → request processing → filter database → content monitoring → content database → content analysis → content formatting → content presentation → **Website/web page**

The Website content management process is classified according to the relevant criteria for forming a response to a user request (Fig. 15):

1. Formation of the content of the web page according to a specific personalized request of the user from the DB at a certain point in time (Fig. 15-Fig. 16). Webpage formation depends on the specific request of each user of the permanent audience. This leads to a significant increase in the load on the Webserver with each user request of a permanent audience of the corresponding Website. The load is reduced by caching frequently requested information in a certain time according to the previous statistical analysis of the dynamics of requests.

2. Formation of static web pages when edited by the Website moderator (Fig. 17-Fig. 18). Full-text content monitoring in large databases/databases is inefficient. The problem of responsiveness and accuracy of content monitoring is solved by IIS in annotated DBs. Effectively apply content monitoring for IIS text according to CSP (templates, annotations) with weighted keywords and stable word combinations with the largest weight values. The problem of the lack of interactive dialogue between the user and the Website is provided not only by the presence of caching of frequently requested data but also by the analysis of the statistics of this client's requests for a certain/entire time.

3. Webpage caching according to the analytics of requests (last similar requests) of users and transitions from IIS with the achievement of visit conversion (Fig. 19-Fig. 21).

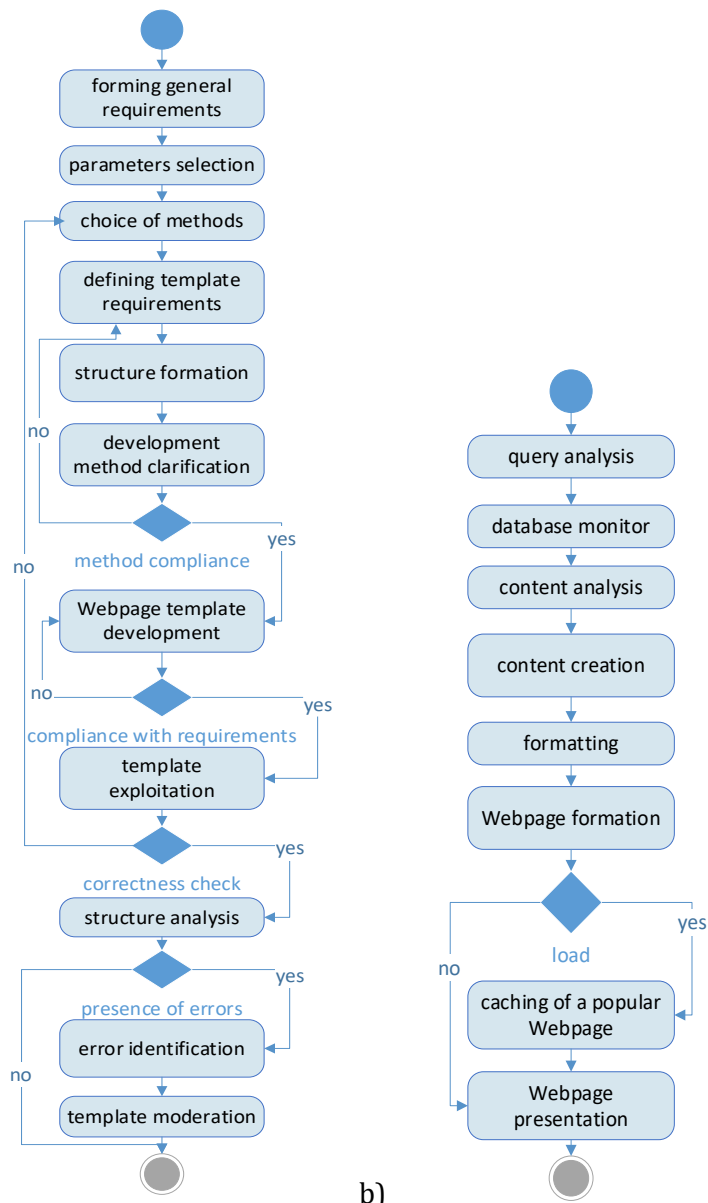


Figure 15: Generation of a) template and b) web page at the request of the user

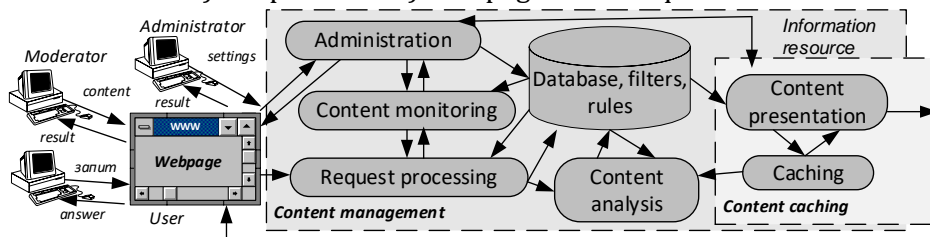


Figure 16: The process of managing content at the user's request

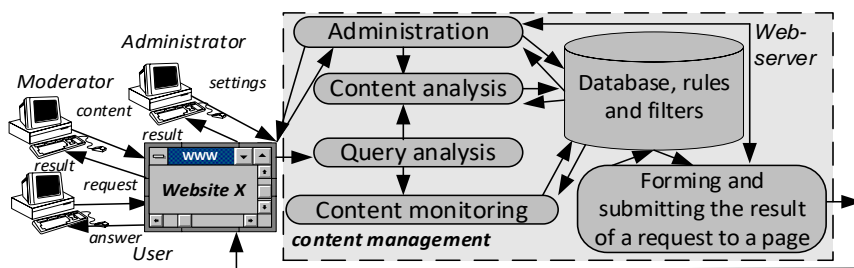


Figure 17: Formation of the static web page during moderation

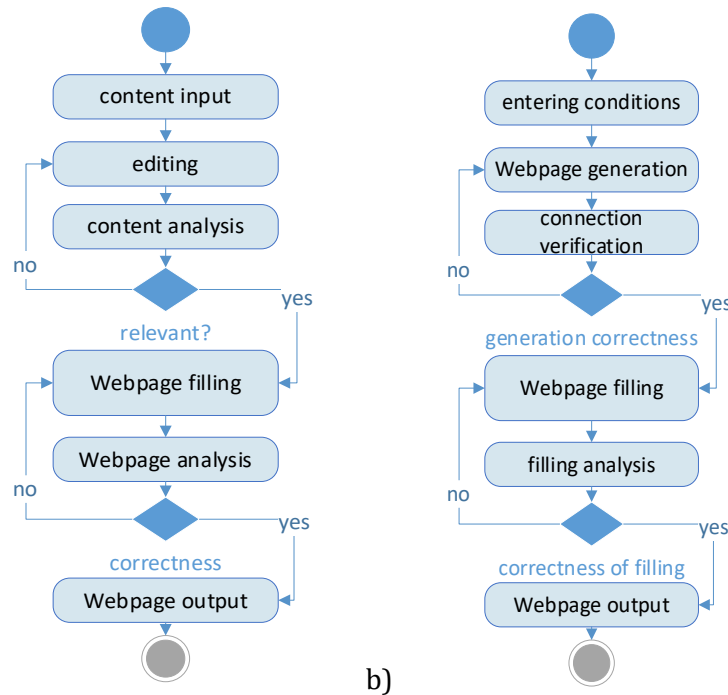


Figure 18: a) Generation and b) filling of the Webpage during content moderation

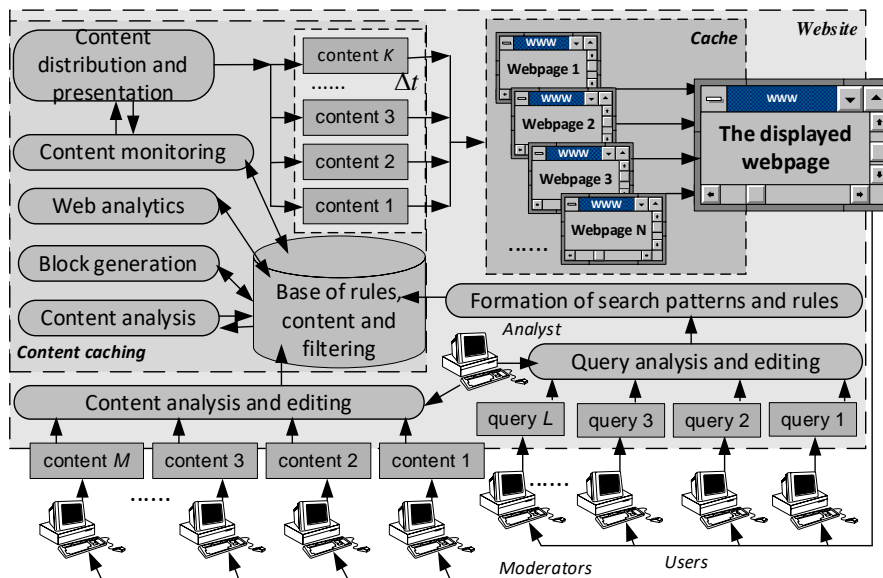


Figure 19: Caching of the generated web page according to query analytics

CLS generates a web page once at a certain moment and stores its image in the DB. The webpage is stored in the cache for a time Δt (as long as the content of the webpage for a certain period Δt is not requested by other users). The set of web pages in the cache is updated according to the history of requests from the permanent audience. Users can access such web pages faster than waiting for a new webpage to be filled. The cache is periodically updated manually/automatically: after the expiration of the term Δt , either the Webpage will not be requested by users for a certain time (Fig. 20), or a significant modification of the Website/content with the content of these Webpages. Analysis of changes in the dynamics and time of access to the relevant cache data determines a set of thematic interests of the permanent audience (Fig. 21). It also determines the speed of development of the needs of end users for the relevant operational thematic areas of the Website content. An appropriate timely analysis of such dynamics of changes in requests and the time of audience interest allows us to adjust the filling of the Website with the relevant content.

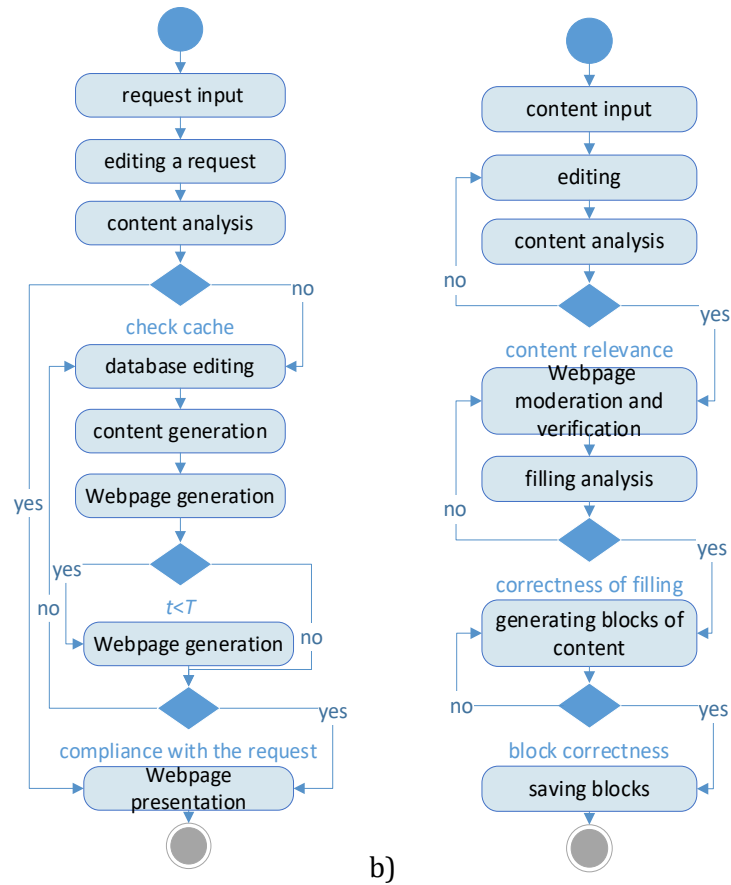


Figure 20: Stages of a) web page generation and b) generation of information blocks for web page generation and caching according to request analytics

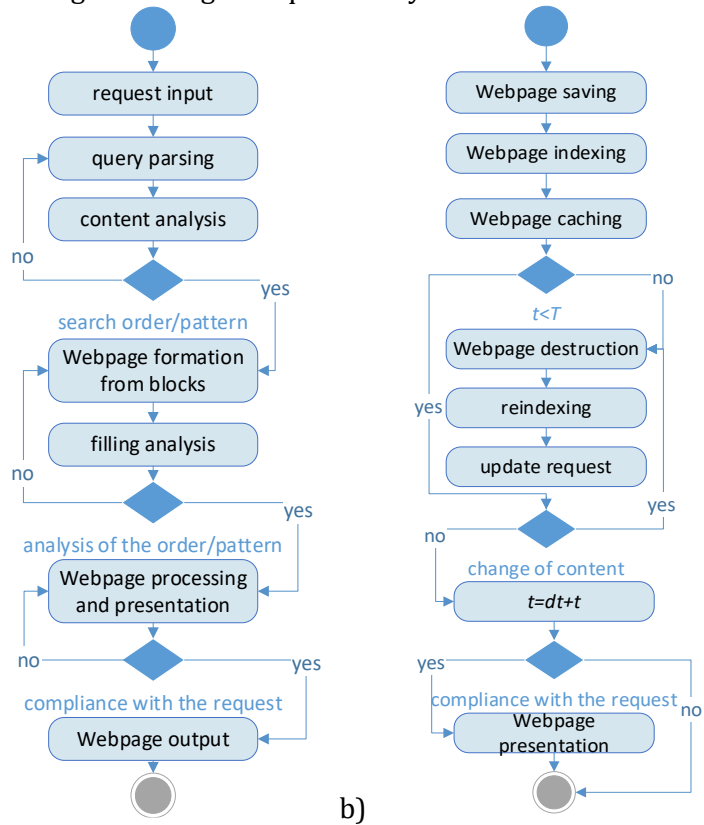


Figure 21: Stages a) presentation of the web page from cached blocks and b) caching of the web page according to the analytics of requests and transitions from IIS

Statistical analysis of links between the content of a text stream allows us to determine the thematic correlation in a certain period and the effectiveness of links to achieve the conversion of user visits (Fig. 22). The use of cluster analysis methods allows you to quantitatively assess the weight of thematic links in text streams in a certain period to predict the popularity of the content topic among each group of regular audiences. This will make it possible to adjust the priority of caching relevant thematic information blocks in a certain period.

The intellectual analysis of the Ukrainian text stream of the Website, taking into account the statistical Web analytics, the achievement of the conversion of user visits is more difficult to successfully implement, taking into account the operational interaction of the permanent user through an interactive flexible dialogue interface to provide access to current relevant content without excess and data noise (Fig. 23).

High-quality, effective, operational and timely analysis of the analytics of requests of regular users, anonymous visitors, conversions from social and IIS by multiple thematic keywords, delay time and actions on a specific target web page, achieving conversions for certain thematic Web pages and rejections for others, etc. will significantly speed up the process analysis of a certain thematic text flow of content for the formation of information blocks, their caching and further content monitoring according to user requests (Fig. 24).

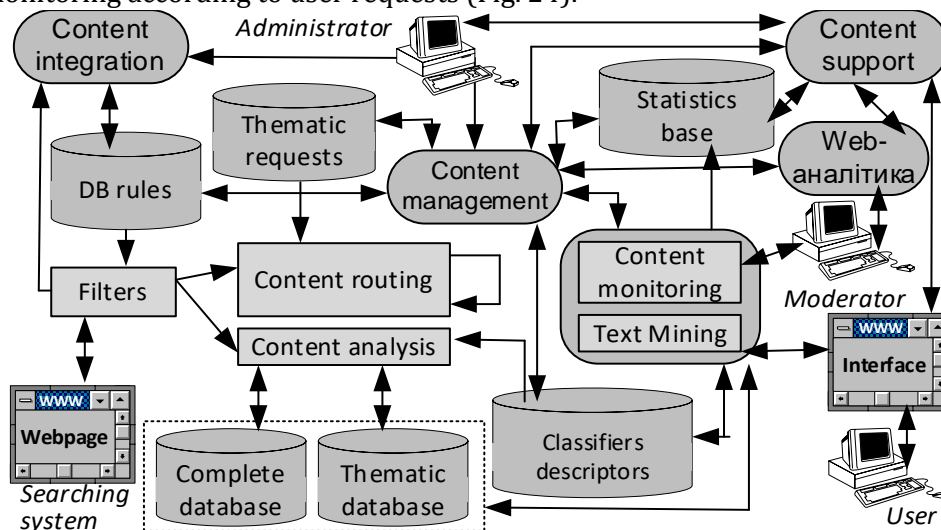


Figure 22: Functional diagram of intellectual text analysis

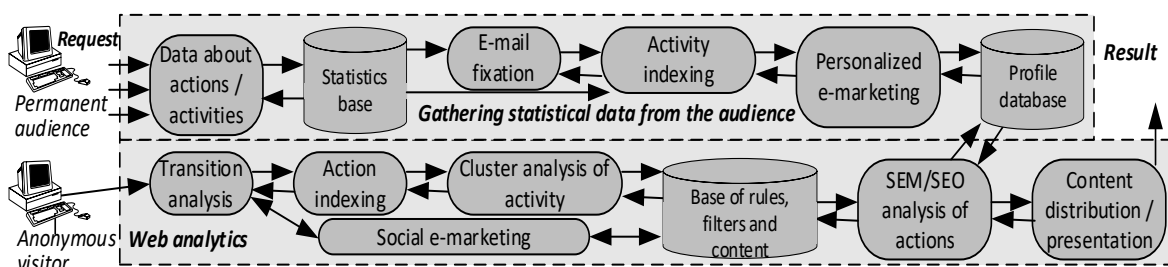


Figure 23: Diagram of the analysis process of user request analytics

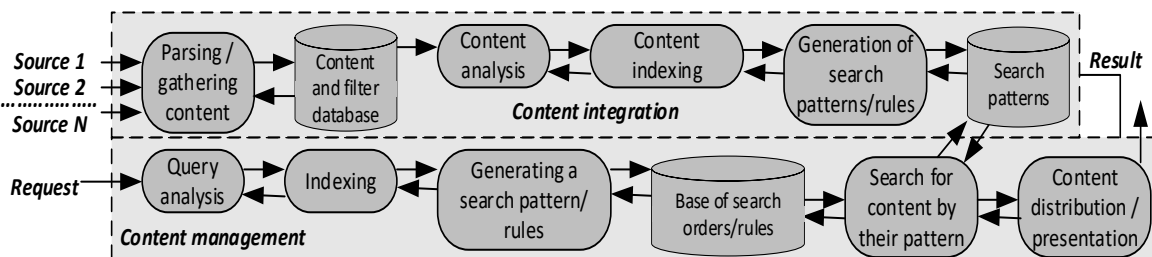


Figure 24: Structural diagram of the text content monitoring process

Significant growth in the amount of content on the Website and variable dynamics, relevance/accuracy/topicalness/timeliness of textual content streams (operational systematic updating) contributes to the growth of content redundancy/noise/sediment, duplication, plagiarism, rewriting and redundancy of IIS requests/results. Integration and content monitoring, content analysis and summarization of a large volume of operational dynamic streams of textual content from Internet sources as a Website requires the implementation of new effective IT IIS/text analytics.

4.3. ML methods of analyzing big data from multiple text content streams

Currently, IT/NLP specialists are actively developing natural language processing software based on machine learning. Some modern ML-based CLSs extract/integrate/generate relevant/relevant/useful content from raw/unstructured information. Such CLS not only analyse natural text but also supports interactive dialogue with the user and adaptation to operational changes in the environment. The results of CLS functioning should be complete/ accurate/ significant, the methods used should be qualitative/effective/intelligent, and accordingly, the IS organization/structure/architecture should be simple/adapted in implementation. These features reveal the underlying methodology for developing CLS based on natural language analysis: clustering similar text into meaningful groups or classifying text based on specific labels, i.e., ML without/with a teacher. A good example is CLS filtering reviews of Yelp Insights [88] based on sentiment analysis, identification of persistent phrases/expressions/phrases, and IIS methods to classify restaurants according to a user's tastes/diets. Another interesting and relevant example is CLS based on accompanying recommendation tags (meta-data about content fragments) implemented by companies such as YouTube, Facebook, Amazon, Netflix, and Stack Overflow. Tags are important for IIS and generating recommendations, as well as in determining the semantic content of content according to the interests of a specific user, etc. Tags identify features of the content they describe, are used for clustering/classification of similar fragments, and offer thematic names for the corresponding clusters.

Google Smart Reply supports the generation of intelligent replies to user e-mails. Voice virtual assistants such as Siri, Alex, Google Assistant and Cortana can analyze speech and give the most likely relevant answers. Siri and Netflix support Ukrainian. Textra, iMessage, and other instant messaging software make predictions about the user's future text based on input, and autocorrect will correct spelling errors. Reverb supports a personalized RSS (news aggregator) based on the Wordnik dictionary. ChatBot Slack accompanies dialogue with context identification.

Linguistic features that make natural language a unique communication tool make it difficult to analyse it based on deterministic rules. The flexibility of human interpretation with more than 50,000 symbolic representations explains the superiority of the average person over any computer in instant understanding of speech. Therefore, fuzzy flexible sensitive computational NLP methods based on machine learning are needed to implement CLS.

The main purpose of ML is to fit existing data to some model of forming a representation of the real world, which helps to make decisions or generate predictions based on new data by finding patterns in it. That is, it is the selection of a set of models to determine the relationships between the target and input data, specifying a shape/pattern with parameters/functions, and minimizing the model error on the input data based on a suitable optimization procedure. The trained model is then fed new data to build a prediction and return markers, probabilities, membership features, or values. The challenge is finding a balance between the ability to find patterns in known data with high accuracy and the ability to generalize to analyse unknown data.

Most natural language analysis software is built on multiple ML models that interact and influence each other. ML models are retrained on new data, using new decision spaces and user-specific tuning to continuously evolve as new content arrives and different aspects of CLS change over time. In CLS, ML models are ranked, aged and deleted (replaced with new ones or modified). That is, CLS ML modules implement content/process life cycles that ensure the correspondence of development dynamics and regional features of natural language with the CLS workflow to support/support/analyse/monitor text content. ML is used to analyse big data from a set of text

streams according to certain characteristics such as diversity, frequency of use, uniqueness, regularity, volume, speed, reliability, time, etc. to solve a specific NLP problem, including error correction. The application of clustering allows you to group linguistic features or typical errors into sets according to corresponding similar characteristics. This is unsupervised ML according to the appropriate algorithm/method: k-means method; DBSCAN (Density-Based Spatial Clustering of Applications with Noise); OPTICS (Ordering Points to Identify the Clustering Structure); PCA (Principal Component Analysis), etc. But the best methods are TF-IDF (Term Frequency – Inverse Document Frequency), singular-value decomposition of the matrix (Singular-Value Decomposition, SVD) and finding cluster groups. Well-known text classification methods are TF-IDF, k-NN method; naive Bayes classifiers, SVM method, latent semantic analysis (LSA), EM algorithm (Expectation-maximization algorithm), decision trees as the ID3/C4.5 algorithm (decision trees), artificial neural networks (ANN), data mining, deep analysis concepts (Concept mining), classification based on Soft set theory or Rough set theory, learning from a set of samples (multiple- instance learning, MIL) and other ML-methods of natural language processing. Recently, deep learning methods have gained popularity.

4.4. Text content clustering with unsupervised ML

In unsupervised ML clustering, the algorithm looks for hidden connections between input data of textual content based on a model of hidden (latent) variables, which includes: the EM algorithm; latent semantic analysis; PCA algorithm; independent component analysis (ICA); BSS method (blind signal separation); the method of moments for finding estimates (Method of moments); non-negative matrix factorization (NMF); hierarchical cluster analysis (HCA) or taxonomy (Taxonomy); singular-value decomposition (SVD), etc. Metrics for the analysis of linguistic units are usually the Rand index, F-measure, Jaccard index, Soren's index (Dice index) and the Fowlkes-Mallows index (Fowlkes-Mallows index) [89]-[99]. The Rand index calculates how similar the clusters are to the reference classifications:

$$I_R = \frac{k_{TP} + k_{TN}}{k_{TP} + k_{TN} + k_{FP} + k_{FN}}, \quad (9)$$

where k_{TP} is the number of true positives, k_{TN} is the number of true negatives; k_{FP} is the number of false positives; k_{FN} is the number of false negatives.

The F-measure is used to balance false negative results by weighting completeness (recall) with the parameter $\varepsilon \geq 0$:

$$I_{FP} = \frac{k_{TP}}{k_{TP} + k_{FP}}, \quad I_{FR} = \frac{k_{TP}}{k_{TP} + k_{FN}}, \quad (10)$$

where I_{FP} is speed of accuracy or precision and I_{FR} is speed of completeness (sensitivity). In IIS:

$$I_{FP} = \frac{|\{a_i\} \cap \{b_j\}|}{|\{b_j\}|}, \quad I_{FR} = \frac{|\{a_i\} \cap \{b_j\}|}{|\{a_i\}|}, \quad (11)$$

where $\{a_i\}$ is the set of relevant content, $\{b_j\}$ is the set of found content. The F-measure is calculated according to the following formula:

$$I_{F\varepsilon} = \frac{(\varepsilon^2 + 1)I_{FP}I_{FR}}{\varepsilon^2 I_{FP} + I_{FR}}, \quad (12)$$

when $\varepsilon = 0$, $I_{F\varepsilon} = I_{FP}$, i.e. I_{FR} does not affect the F-measure of $I_{F\varepsilon}$ at $\varepsilon = 0$, and increasing ε assigns an increasing amount of weight to I_{FR} in the final F-measure.

The Jaccard index is used to quantify the similarity between two data sets X and Y (takes values from 0 to 1) and is defined as:

$$I_{Jcr} = \frac{|X \cap Y|}{|X \cup Y|} = \frac{k_{TP}}{k_{TP} + k_{FP} + k_{FN}}, \quad (13)$$

It is the number of unique elements common to both sets divided by the total number of unique elements in both sets. An index of 1 means that two data sets are identical, and an index of 0 indicates that the data sets have no common elements.

The Sorens index doubles the weight of k_{TP} while ignoring k_{TN} :

$$I_{DSC} = \frac{2k_{TP}}{2k_{TP} + k_{FP} + k_{FN}}. \quad (14)$$

The Fowlkes-Mallows index calculates the similarity between the returned clusters and the reference classifications. The higher the value of I_{FM} , the more similar the clusters and reference classifications are:

$$I_{FM} = \sqrt{\frac{k_{TP}}{k_{TP} + k_{FP}} \frac{k_{TP}}{k_{TP} + k_{FN}}}, \quad (15)$$

where k_{TP} is the number of true positive results, k_{FP} is the number of false positives, k_{FN} is the number of false negatives. The I_{FM} index is the geometric mean of the precision of I_{FP} and the completeness of I_{FR} , and is known as the G-measure, and the F-measure is their harmonic value. In addition, I_{FP} and I_{FR} are known as Wallace's indices I' and I'' . The normalized samples I_{FP} , I_{FR} and G-measures correspond to the information index I_{YJS} (Youden's index or Youden's J statistic), the markedness index I_M , the Matthews correlation coefficient I_{MCC} (Matthews correlation coefficient (MCC) or phi coefficient) and strongly related to Cohen's kappa coefficient $I_{CK\kappa}$ (English Cohen's kappa coefficient, κ). The Wallace index I_{YJS} captures the effectiveness of a dichotomous diagnostic experiment through sensitivity and specificity analysis:

$$I_{YJS} = \frac{k_{TP}}{k_{TP} + k_{FN}} + \frac{k_{TN}}{k_{TN} + k_{FP}} - 1. \quad (16)$$

Informedness is a generalization of I_{YJS} to the multi-class case and estimates the probability of an informed decision. The Matthews correlation coefficient I_{MCC} is calculated as the Pearson phi coefficient:

$$I_{MCC} = \frac{k_{TP}k_{TN} - k_{FP}k_{FN}}{\sqrt{(k_{TP} + k_{FP})(k_{TP} + k_{FN})(k_{TN} + k_{FP})(k_{TN} + k_{FN})}}. \quad (17)$$

Cohen's Kappa coefficient $I_{CK\kappa}$ is a measure of inter-rater reliability and intra-rater reliability for qualitative/categorical items:

$$I_{CK\kappa} = \frac{2(k_{TP}k_{TN} - k_{FP}k_{FN})}{(k_{TP} + k_{FP})(k_{FP} + k_{TN}) + (k_{TP} + k_{FN})(k_{FN} + k_{TN})}. \quad (18)$$

There is controversy surrounding Cohen's Kappa coefficient $I_{CK\kappa}$ due to difficulties in interpreting the consistency indicators. Some researchers suggest that it is conceptually easier to estimate differences between elements.

4.5. Main areas of research

Today, there are many computer linguistic systems for various purposes, even for processing Ukrainian-language textual content. But these are usually commercial projects of a closed type (there are no publications or access to the administrative part) and most often they are foreign projects. There seem to be a lot of publications to understand how the natural language processing process generally works, especially for English texts. However, applying these models, methods, algorithms and technologies directly to Ukrainian-language textual content does not lead to almost any positive result. Already at the level of morphological analysis, a significant conflict arises between the developed methods and the incoming Ukrainian text - the output is not correct. For example, for a simple Porter algorithm (stemming) without a corresponding modification, it will not be correct to separate the base of the word from the inflexion, which will lead to incorrect identification of the keywords of the texts, which in turn affects any NLP task where it is necessary to quickly identify a set of keywords (rubrication, search, annotation, etc.). Determining the main processes and features of the linguistic analysis of Ukrainian-language texts will greatly facilitate the stages of processing the text flow of content such as integration, support and content management (Fig. 25). In turn, the adaptation of the processes of intellectual analysis of text content with the identification of functional requirements for the corresponding CLS modules will lead to the possibility of developing a typical architecture of similar systems based on the principle of modularity (adding components depending on the content of the NLP task and the purpose of the CLS).

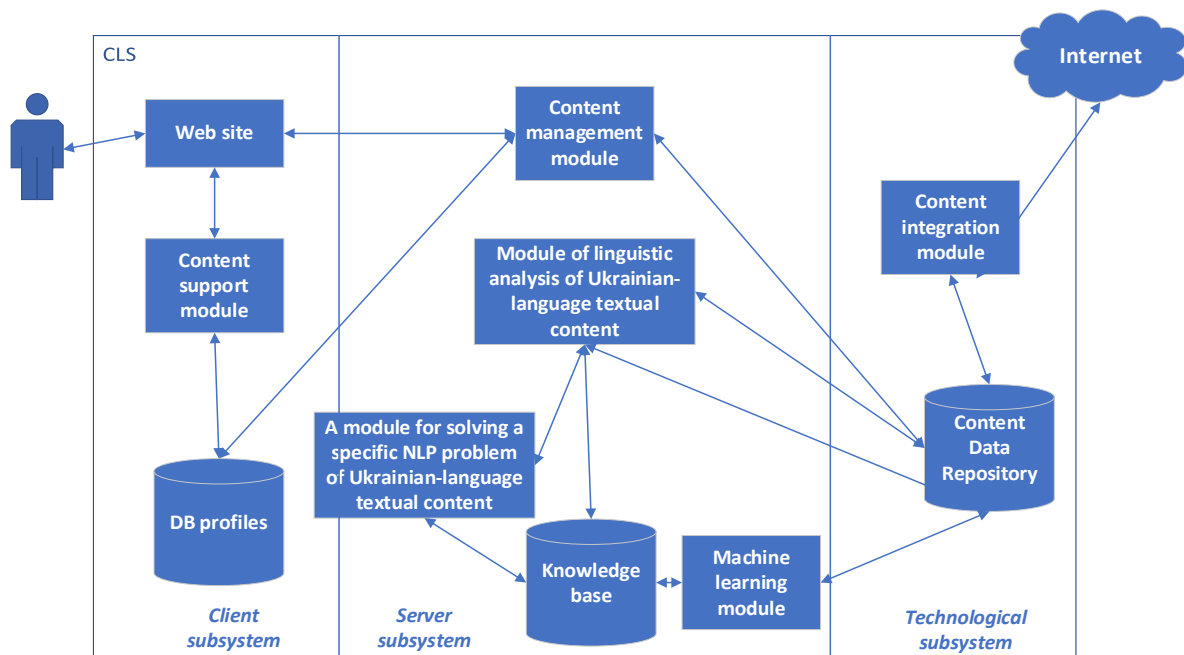


Figure 25: Generalized architecture of a computer linguistic system

The application of the specified technologies/methods/models in a typical CLS architecture, adapted for any NLP task of processing Ukrainian-language textual content, is a necessary prerequisite for the successful implementation of a computer linguistic system project for solving a specific NLP task that requires the use of appropriate set standard libraries, utilities and open-source software that will solve specialized project tasks according to the end user's needs.

5. Conclusions

The analysis and synthesis of CLS is based on the application of linguistic analysis of Ukrainian-language textual content, intelligent processing of the textual flow of content, machine learning of the system based on reliable data, and statistical analysis to find patterns in the appearance of linguistic events. An analysis of the current state and prospects for IT development of natural language processing was carried out, which made it possible to define the problem and research tasks, as well as to form general research directions in the absence of non-commercial CLS with open source for processing Ukrainian-language textual content and a standardized design approach. The concept of CLS is defined and their general classification is given. A detailed analysis of the known CLS was carried out, which made it possible to improve the general classification of the corresponding IS. The main NLP tasks of computer linguistic systems are defined, based on which examples and a comparative analysis of known modern CLS are given. This made it possible to form general directions of research. The main general scheme of the process of linguistic analysis of text in natural language using CLS tools is described and analyzed. The main states and properties of CLS, their classification and features are defined. Well-known classical approaches and trends in natural language processing are analysed. A general classification of the main NLP approaches, directions and additional methods of linguistic research for NLP tasks is presented. An analysis of the existing basic methods and methods of processing natural language using machine learning was carried out. Their classification was carried out and typical problems of ML-methods for processing Ukrainian-language texts were determined. An overview of the known IT development of CLS based on the features and technologies of intellectual analysis of the flow of Ukrainian-language content was made. The main requirements for CLS performance evaluation based on ML technology and big data analysis are defined. Basic ML for analysing big data from multiple textual content streams is reviewed. Requirements for text content clustering in unsupervised ML are defined.

References

- [1] The free dictionary by Farlex. Linguistic System. URL: <https://encyclopedia2.thefreedictionary.com/Linguistic+System>.
- [2] Glottopedia. Linguistic information system. URL: http://www.glottopedia.org/index.php/Linguistic_information_system.
- [3] Lenhart Schubert. Computational linguistics. Stanford Encyclopedia of Philosophy. URL: <https://plato.stanford.edu/entries/computational-linguistics/>.
- [4] R. Nazarchuk, S. Albota, Tweets about Ukraine during the russian-Ukrainian War: Quantitative Characteristics and Sentiment Analysis, CEUR Workshop Proceedings 3426 (2023) 551-560.
- [5] R. Romanchuk, V. Vysotska, V. Andrunyk, L. Chyrun, S. Chyrun, O. Brodyak, Intellectual Analysis System Project for Ukrainian-language Artistic Works to Determine the Text Authorship Attribution Probability, in: Proceedings of the 18th IEEE International Conference on Computer Science and Information Technologies, CSIT 2023, Lviv, Ukraine, October 19-21, 2023. IEEE 2023.
- [6] M. Konyk, V. Vysotska, S. Goloshchuk, R. Holoshchuk, S. Chyrun, I. Budz, Technology of Ukrainian-English Machine Translation Based on Recursive Neural Network as LSTM, CEUR Workshop Proceedings 3387 (2023) 357-370.
- [7] V. Vysotska, Y. Burov, V. Lytvyn, A. Demchuk, Defining Author's Style for Plagiarism Detection in Academic Environment, in: Proceedings of the International Conference on Data Stream Mining and Processing, DSMP, 2018, pp. 128-133. doi: 10.1109/DSMP.2018.8478574.
- [8] A. Berko, Y. Matseliukh, Y. Ivaniv, L. Chyrun, V. Schuchmann, The text classification based on Big Data analysis for keyword definition using stemming, in: Proceedings of the IEEE 16th International conference on computer science and information technologies, CSIT-2021, Lviv, Ukraine, 22-25 September 2021, pp. 184-188.
- [9] V. Lytvyn, V. Vysotska, I. Budz, Y. Pelekh, N. Sokulska, R. Kovalchuk, L. Dzyubyk, O. Tereshchuk, M. Komar, Development of the quantitative method for automated text content authorship attribution based on the statistical analysis of N-grams distribution, Eastern-European Journal of Enterprise Technologies 6(2(102)) (2019) 28-51. doi: 10.15587/1729-4061.2019.186834.
- [10] R. Lynnyk, V. Vysotska, Y. Matseliukh, Y. Burov, L. Demkiv, A. Zaverbnyj, A. Sachenko, I. Shylinska, I. Yevseyeva, O. Bihun, DDOS attacks analysis based on machine learning in challenges of global changes, CEUR Workshop Proceedings 2631 (2020) 159-171.
- [11] T. Batura, A. Bakiyeva, M. Charintseva, A method for automatic text summarization based on rhetorical analysis and topic modeling, International Journal of Computing 19(1) (2020) 118-127. doi: 10.47839/ijc.19.1.1700.
- [12] N. Shakhovska, I. Shvorob, The method for detecting plagiarism in a collection of documents, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2015, pp. 142-145.
- [13] O. Barkovska, V. Kholiev, A. Havrashenko, D. Mohylevskyi, A. Kovalenko, A Conceptual Text Classification Model Based on Two-Factor Selection of Significant Words, CEUR Workshop Proceedings 3396 (2023) 244-255.
- [14] I. Khomytska, I. Bazylevych, V. Teslyuk, I. Karamysheva, The chi-square test and data clustering combined for author identification, in: Proceedings of the IEEE XVIIIth Scientific and Technical Conference on Computer Science and Information Technologies, CSIT 2023, Lviv, Ukraine, 19-21 October 2023.
- [15] I. Khomytska, V. Teslyuk, The Multifactor Method Applied for Authorship Attribution on the Phonological Level, CEUR workshop proceedings 2604 (2020) 189-198.
- [16] I. Khomytska, V. Teslyuk, A. Holovatyy, O. Morushko, Development of methods, models, and means for the author attribution of a text, Eastern-European Journal of Enterprise Technologies. 3(2(93)) (2018) 41-46. doi: 10.15587/1729-4061.2018.132052.

- [17] I. Khomytska, V. Teslyuk, Authorship and Style Attribution by Statistical Methods of Style Differentiation on the Phonological Level, *Advances in Intelligent Systems and Computing* 871 (2019) 105–118. doi: 10.1007/978-3-030-01069-0_8.
- [18] A. Taran, Terminology of Computational Linguistics in Terms of Indexing and Information Retrieval in the System "iSybislaw", *CEUR Workshop Proceedings* 2870 (2021) 225-234.
- [19] N. Kunanets, H. Matsiuk, Use of the Smart City Ontology for Relevant Information Retrieval, *CEUR Workshop Proceedings* 2362 (2019) 322-333.
- [20] K. Nataliia, M. Halyna, Application of Saaty Method While Choosing Thesaurus View Model of the "Smart city" Subject Domain for the Improvement of Information Retrieval Efficiency, in: *Proceedings of the IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2018, Vol. 2, Art No. 8526656, 2018*, pp. 21-25. doi: 10.1109/STC-CSIT.2018.8526656.
- [21] E. Fedorov, O. Nechyporenko, Linguistic Constructions Translation Method Based on Neural Networks, *CEUR Workshop Proceedings* 3396 (2023) 295-306.
- [22] V. Lytvyn, Y. Burov, V. Vysotska, Y. Pukach, O. Tereshchuk, I. Shakleina, Abstracting Text Content Based on Weighing the TF-IDF Measure by the Subject Area Ontology, in: *Proceedings of the IEEE International Conference on Smart Information Systems and Technologies (SIST), Nur-Sultan, Kazakhstan, 2021*. URL: <https://ieeexplore.ieee.org/document/9465978>.
- [23] O. Bisikalo, O. Boivan, N. Khairova, O. Kovtun, V. Kovtun, Precision automated phonetic analysis of speech signals for information technology of text-dependent authentication of a person by voice, *CEUR Workshop Proceedings* 2853 (2021) 276–288.
- [24] A. Sartiukova, O. Markiv, V. Vysotska, I. Shakleina, N. Sokulska, I. Romanets, Remote Voice Control of Computer Based on 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. 1058-1064.
- [25] 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.
- [26] 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.
- [27] N. Shakhovska, O. Basystiuk, K. Shakhovska, Development of the Speech-to-Text Chatbot Interface Based on Google API, *CEUR Workshop Proceedings* 2386 (2019) 212-221.
- [28] V. Lytvyn, V. Vysotska, P. Pukach, Z. Nytrebych, I. Demkiv, R. Kovalchuk, N. Huzyk, Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients, *Eastern-European Journal of Enterprise Technologies* 5 (2(95)) (2018) 16–28. doi: 10.15587/1729-4061.2018.142451.
- [29] V. Lytvyn, V. Vysotska, V. Kuchkovskiy, I. Bobyk, O. Malanchuk, Y. Ryshkovets, et. al. Development of the system to integrate and generate content considering the cryptocurrent needs of users, *Eastern-European Journal of Enterprise Technologies* 1(2(97)) (2019) 18–39. doi: 10.15587/1729-4061.2019.154709
- [30] A. Chiche, H. Kadi, T. Bekele, A Hidden Markov Model-based Part of Speech Tagger for Shekki'noono Language, *International Journal of Computing* 20(4) (2021) 587-595. doi: 10.47839/ijc.20.4.2448.
- [31] S. A. Thorat, K. P. Jadhav, Improving Conversation Modelling using Attention Based Variational Hierarchical RNN, *International Journal of Computing* 20(1) (2021) 39-45. doi: 10.47839/ijc.20.1.2090.
- [32] I. Lauriola, A. Lavelli, F. Aiolli, An introduction to deep learning in natural language processing: Models, techniques, and tools, *Neurocomputing* 470 (2022) 443-456.
- [33] Y. Kang, et. al., Natural language processing (NLP) in management research: A literature review, *Journal of Management Analytics* 7(2) (2020) 139-172.

- [34] L. Hickman, S. Thapa, L. Tay, M. Cao, P. Srinivasan, Text preprocessing for text mining in organizational research: Review and recommendations, *Organizational Research Methods* 25(1) (2022) 114-146.
- [35] D. Hu, An introductory survey on attention mechanisms in NLP problems, in: *Proceedings of the 2019 Intelligent Systems Conference (IntelliSys)*, Volume 2, 2020, pp. 432-448.
- [36] Gardner, M., Merrill, W., Dodge, J., Peters, M. E., Ross, A., Singh, S., & Smith, N. A. (2021). Competency problems: On finding and removing artifacts in language data. arXiv preprint arXiv:2104.08646.
- [37] L. Wu, et al., Graph neural networks for natural language processing: A survey, *Foundations and Trends® in Machine Learning* 16(2) (2023). 119-328.
- [38] M.-A. Lefer, N. Grabar, Super-creative and overbureaucratic: A cross-genre corpusbased study on the use and translation of evaluative prefixation in ted talks and eu parliamentary debates, *Across Languages and Cultures* 16(2) (2015) 187-208.
- [39] D. Jurafsky, J. H. Martin, *Speech and Language Processing*. URL: https://web.stanford.edu/~jurafsky/slp3/ed3book_sep212021.pdf.
- [40] J. Weizenbaum, ELIZA – A computer program for the study of natural language communication between man and machine, *CACM* 9 (1966) 36-45.
- [41] J. Weizenbaum, *Computer Power and Human Reason: From Judgement to Calculation*. W.H. Freeman and Company. 1976.
- [42] ElizaBot. URL: <https://www.masswerk.at/elizabot/>.
- [43] ELIZA: a very basic Rogerian psychotherapist chatbot. URL: <https://web.njit.edu/~ronkowitz/eliza.html>.
- [44] D. Jurafsky, J. H. Martin, Regular Expressions, Text Normalization, Edit Distance. URL: <https://web.stanford.edu/~jurafsky/slp3/2.pdf>.
- [45] O. Karnalim, G. Kurniawati, Programming style on source code plagiarism and collusion detection, *International Journal of Computing* 19(1) (2020) 27-38.
- [46] V. Claveau, T. Hamon, S. Le Maguer, N. Grabar, Health consumer-oriented information retrieval, *Studies in Health Technology and Informatics* 210 (2015) 80-84.
- [47] P. Zweigenbaum, S.J. Darmoni, N. Grabar, The contribution of morphological knowledge to French MeSH mapping for information retrieval, in: *Proceedings of the AMIA Symposium*, 2001, pp. 796-800.
- [48] É. Bigeard, F. Thiessard, N. Grabar, Detecting drug non-compliance in internet fora using information retrieval and machine learning approaches, *Studies in Health Technology and Informatics* 264 (2019) 30-34.
- [49] V. Claveau, T. Hamon, S. Le Maguer, N. Grabar, Health consumer-oriented information retrieval, *Studies in Health Technology and Informatics* 210 (2015) 80-84.
- [50] A. Périnet, T. Hamon, Distributional analysis applied to specialized texts. Reduction of data sparseness by context abstractions, *TAL Traitement Automatique des Langues* 56(2) (2015) 77-102.
- [51] V. Trysnyuk, Y. Nagorny, K. Smetanin, I. Humeniuk, T. Uvarova, A method for user authenticating to critical infrastructure objects based on voice message identification, *Advanced Information Systems* 4(3) (2020) 11-16. doi: 10.20998/2522-9052.2020.3.02.
- [52] A. Medvedyk, M. Lohoida, Z. Rybchak, O. Kulyna, IT Slang: Development of Telegram Chatbot, *CEUR Workshop Proceedings* 3396 (2023) 152-162.
- [53] O. Romanovskiy, et al., Elomia Chatbot: The Effectiveness of Artificial Intelligence in the Fight for Mental Health, *CEUR Workshop Proceedings* 2870 (2021) 1215-1224.
- [54] A. Yarovy, D. Kudriavtsev, Method of Multi-Purpose Text Analysis Based on a Combination of Knowledge Bases for Intelligent Chatbot, *CEUR Workshop Proceedings* 2870 (2021) 1238-1248.
- [55] 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.

- [56] 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.
- [57] 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.
- [58] S. Yevseiev, et al., Development of a method for determining the indicators of manipulation based on morphological synthesis, Eastern-European Journal of Enterprise Technologies 117(9) (2022) 22-35.
- [59] O. Cherednichenko, O. Kanishcheva, O. Yakovleva, D. Arkatov, Collection and Processing of a Medical Corpus in Ukrainian, CEUR Workshop Proceedings 2604 (2020) 272-282.
- [60] A. Dmytriv, V. Vysotska, M. Bublyk, The Speech Parts Identification for Ukrainian Words Based on VESUM and Horokh Using, in: International Conference on Computer Sciences and Information Technologies, CSIT-2021, September 2021, Vol. 2, pp. 21-33.
- [61] 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: Computer Sciences and Information Technologies, CSIT-2022, November 2022, pp. 93-98.
- [62] M. Lupei, et al., Analyzing Ukrainian Media Texts by Means of Support Vector Machines: Aspects of Language and Copyright, in: Computer Science, Engineering and Education Applications, 2023, March, pp. 173-182. Cham: Springer Nature Switzerland.
- [63] V. Vysotska, Analytical Method for Social Network User Profile Textual Content Monitoring Based on the Key Performance Indicators of the Web Page and Posts Analysis, CEUR Workshop Proceedings 3171 (2022) 1380-1402.
- [64] K. Shakhovska, et al., An approach for a next-word prediction for Ukrainian language. Wireless Communications and Mobile Computing 2021 (2021) 1-9.
- [65] P. Zhezhnych, A. Shilinh, I. Demydov, Architecture of the Computer-linguistic System for Processing of Specialized Web-communities' Educational Content, CEUR Workshop Proceedings 2616 (2020) 1-11.
- [66] V. Vysotska Ukrainian participles formation by the generative grammars use, CEUR Workshop Proceedings 2604 (2020) 407-427.
- [67] B. Bengfort, R. Bilbro, T. Ojeda, Applied text analysis with Python: Enabling language-aware data products with machine learning. O'Reilly Media, Inc. (2018).
- [68] D. Jurafsky, J. H. Martin, Deep Learning Architectures for Sequence Processing. URL: <https://web.stanford.edu/~jurafsky/slp3/9.pdf>.
- [69] D. Jurafsky, J. H. Martin, Naive Bayes and Sentiment Classification. URL: <https://web.stanford.edu/~jurafsky/slp3/4.pdf>.
- [70] D. Jurafsky, Logistic Regression. URL: <https://web.stanford.edu/~jurafsky/slp3/5.pdf>.
- [71] D. Jurafsky, J. H. Martin, Neural Networks and Neural Language Models. <https://web.stanford.edu/~jurafsky/slp3/7.pdf>.
- [72] P. Kravets, The Game Method for Orthonormal Systems Construction, in: Proceeding of the 9th International Conference - The Experience of Designing and Applications of CAD Systems in Microelectronics, 2007. doi: 10.1109/cadsm.2007.4297555.
- [73] M. Johnson, G. Lakoff, Why cognitive linguistics requires embodied realism, Cognitive Linguistics, 2002. doi: 10.1515/cogl.2002.016.
- [74] M. Rehani, W. L. Wolf, Methods and systems for measuring semantics in communications. <https://patentimages.storage.googleapis.com/00/d2/da/886c00fc2dce4b/US9269353.pdf>.
- [75] L. A. Kovbasyuk, I. O. Fritsky, V. N. Kokozay, T. S. Iskenderov, Synthesis and structure of diaqua-bis (ethylenediamine) copper (II) salts with anions of carbamic acids, Polyhedron 16(10) (1997) 1723-1729.
- [76] L. A. Kovbasyuk, O. A. Babich, V. N. Kokozay, Direct synthesis and crystal structure of a mixed-valence copper complex, Polyhedron 16(1) (1997) 161-163.
- [77] D. Lande, L. Strashnoy, GPT Semantic Networking: A Dream of the Semantic Web-The Time is Now. URL: <https://ela.kpi.ua/server/api/core/bitstreams/299901e4-b9b9-457b-9f07-a0808f3973ba/content>.

- [78] D. Lande, et al., Link prediction of scientific collaboration networks based on information retrieval, *World Wide Web* 23 (2020) 2239-2257.
- [79] M. Fu, et al. Integration of complete ensemble empirical mode decomposition with deep long short-term memory model for particulate matter concentration prediction, *Environmental Science and Pollution Research* 28 (2021) 64818-64829.
- [80] M. Fu, J. Feng, D. Lande, O. Dmytrenko, D. Manko, R. Prapakovich, Dynamic model with super spreaders and lurker users for preferential information propagation analysis, *Physica A: statistical mechanics and its applications* 561 (2021) 125266.
- [81] D. V. Lande, A. A. Snarskii, E. V. Yagunova, E. V. Pronoza, The use of horizontal visibility graphs to identify the words that define the informational structure of a text, in: *IEEE 12th Mexican International Conference on Artificial Intelligence*, 2013, November, pp. 209-215.
- [82] O. Oborska, M. Teliatynskiy, D. Dosyn, V. Lytvyn, S. Kostenko, An Intelligent System Based on Ontologies for Determining the Similarity of User Preferences, *CEUR Workshop Proceedings* 3403 (2023) 283-292.
- [83] D. Dosyn, Y. I. Daradkeh, V. Kovalevych, M. Luchkevych, Y. Kis, Domain Ontology Learning using Link Grammar Parser and WordNet, *CEUR Workshop Proceedings* 3312 (2022) 14-36.
- [84] Y. Burov, K. Mykich, I. Karpov, Intelligent systems based on ontology representation transformations, in: *Conference on Computer Science and Information Technologies*, 2020, September, pp. 263-275. Cham: Springer International Publishing.
- [85] Y. Burov, Knowledge Based Situation Awareness Process Based on Ontologies, *CEUR Workshop Proceedings* 2870 (2021) 413-423.
- [86] Y. Burov, K. Mykich, I. Karpov, Building a versatile knowledge-based system based on reasoning services and ontology representation transformations, in: *IEEE 15th International Conference on Computer Sciences and Information Technologies*, 2020, pp. 255-260.
- [87] B. Clifton, *Advanced web metrics with Google Analytics*. John Wiley & Sons. 2012.
- [88] Yelp Insights. URL: <https://blog.yelp.com/news/yelpy-insights/>.
- [89] R. Anita, C. N. Subalalitha, An approach to cluster Tamil literatures using discourse connectives, in: *IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, 2019, pp. 1-4.
- [90] O. Tverdokhlib, V. Vysotska, P. Pukach, M. Vovk, Information Technology for Identifying Hate Speech in Online Communication Based on Machine Learning, *Data-Centric Business and Applications: Modern Trends in Financial and Innovation Data Processes* 1 (2024) 339-369.
- [91] D. Nakache, E. Metais, J. F. Timsit, Evaluation and NLP, in: *International Conference on Database and Expert Systems Applications*. Springer, Berlin, Heidelberg, 2005, pp. 626-632.
- [92] M. Tikhonova, A. Gavrishchuk, NLP methods for automatic candidate's cv segmentation, in: *IEEE International Conference on Engineering and Telecommunication*, 2019. pp. 1-5.
- [93] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, J. Li, Dice loss for data-imbalanced NLP tasks, *arXiv preprint 2019*. arXiv:1911.02855.
- [94] Ryu Keun Ho, BioBERT Based Efficient Clustering Framework for Biomedical Document Analysis, in: *Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computing*, October 21-23, 2021, Jilin, China. Springer Nature. p. 179.
- [95] N. Rayzmann, H. Aponso, C. Y. Markgraf, P. E. Chappell, SUN-238 Estrogen Modulates Expression Levels of Gonadotropin-Releasing Hormone Receptor (GNRHR) in Immortalized Kisspeptin Neurons in Vitro, *Journal of the Endocrine Society* 4 (2020) SUN-238.
- [96] Y. Tan, et al., Triaging ophthalmology outpatient referrals with machine learning: a pilot study, *Clinical & experimental ophthalmology* 48(2) (2020) 169-173.
- [97] Kim Ju-Ri, Using Markedness Principle for Abstraction of Dependency Relations of Natural Languages, *Eurasian Journal of Applied Linguistics* 7.2 (2021) 58-67.
- [98] D. Heo, W. Lee, B. Jung, J. H. Lee, Quality estimation using dual encoders with transfer learning, in: *Proceedings of the Sixth Conference on Machine Translation*, 2021, pp. 920-927.
- [99] K. Ayre, et al., Developing a natural language processing tool to identify perinatal self-harm in electronic healthcare records, *PloS one* 16(8) (2021) e0253809.