Stepanova Iryna Serhiivna

Candidate of Philology, Associate Professor,
Head of the Department of Foreign Languages
Vinnytsia national technical university

Nykyporets Svitlana Stepanivna

Lecturer of the Department of Foreign Languages
Vinnytsia national technical university
ORCID ID: 0000-0002-3546-1734

THE INTERNET AND LINGUISTICS: INTERACTION AND NEW PROSPECTS OF CORPUS RESEARCH

Annotation. *An effort to analyse the linguistic research of the Internet discourse is made in the article. The authors believe that the creation of linguistic corpora of the Ukrainian language at the present stage is not systemic. The authors also consider prospects and possible approaches to the Internet text space by means of Corpus Linguistics – a fairly new field of linguistics, closely related to computational and cognitive linguistics.*

Key words**:** *language corpus, Internet discourse, corpus linguistics, foreign language teaching methodology, linguistic paradigm.*

Launched as a military project [1], the Internet has rapidly evolved from a data channel into a medium of communication, which in turn has led to a global information environment, a large and diverse Internet world, which is the subject of many humanities disciplines, including linguistics.

Anthropocentrism as a leading general scientific paradigm of the XXI century manifests itself in such aspects as linguistic research of Internet communication, computer-mediated communication and the language of the Internet, learning the specifics of speech, genres, vocabulary and more. [2] On the other hand, the Internet is a virtually inexhaustible source of authentic texts, making it the subject of corpus linguistics.

Psychological and sociolinguistic research of the Internet includes the study of such aspects as gender, identification and presentation of the virtual personality, the communicative space of the Web as an environment for the functioning of computer (or electronic) discourse. Studies by L. F. Kompantseva, O. I. Goroshko, O. E. Voiskunsky and others were performed in this direction. Among foreign authors it is worth mentioning G. Lerner. E. Sherman, H. Clark, M. Costels. The researches of L. U. Ivanov, S. A. Nedobukh, G. M. Trohymova, O. M. Galichkina, A. E. Zhychkina, as well as J. Barbatis, J. Nilson, R. Duokins, D. Crystal, R. Dixon and others. The interest of researchers in the functional and stylistic characteristics of Internet communication, highlighting the specifics of genres is reflected in the works of V. P. Zakharov, O. V. Buldakov, L. A. Kapanadze, P. Lynch, S. Horton, J. Challenger.

The works of M. Folk, G. Greffenstet, F. Resnik, A. Kilgarif, E. Aguirre, D. Martinez are devoted to the use of the Internet as a linguistic corpus. The works of these authors study both idiolects, in particular, authorial, and generalize the methodology that forms the content of corpus linguistics, theory and methodology, the creation of corpora, as well as the actual corpus research, i.e. the study of various aspects of language using corpus methods. The creation of linguistic corpora of the Ukrainian language at the present stage is not systemic.

Ukrainian linguistic research of the Internet today is mainly conducted in cognitive, linguopragmatic, linguocultural viewpoints (L. F. Kompantseva, O. V. Dmytruk, O. V. Vinareva, M. O. Stolyarova and others) mainly on English language material.

The purpose of our article is to analyse the mutual influence of traditional and computational linguistics, to determine the role of the Internet in the development of corpus linguistics, its promising areas, including methodological and didactic.

Despite the fact that F. de Saussure’s statement about the dualism of the nature of language, his langue-parole dichotomy is generally accepted in science, it cannot be stated unequivocally that modern linguistics has one convincing theory that explains the asymmetric nature of language embodied in the thesis of S. Kartsevsky.

What ultimately is the object of linguistic research: langue or parole, language or speech, paradigm or syntagm? There is no consensus on this issue, but it is clear that a combination of intuitive and textual approaches can ensure the verifiability of scientific results. The combination of traditional linguistic research approaches with modern information technologies has made it possible to create extremely large in size and diverse in nature sets of language and speech material. These sets are called corpora, and the field of linguistics that uses such tools is called corpus linguistics. [3]

Of course, the first corpora appeared long before the advent of electronic computing methods. Diachronic research has always been based on the introspection of some limited textual material. Dialectological field excursus also summarize a certain set of natural speech material. The transition to electronic forms of storage and processing of linguistic corpora does not change the essence of the research methodology, but infinitely expands its capabilities and prospects. Thus, corpus linguistics is a fairly new field of linguistics, closely related to computational and cognitive linguistics. It is united by the technology and tools of processing language (text) material with the first, and it is common with the second in the basic postulate: the object of its interest is speech activity, represented by an infinite and inexhaustible number of texts. Corpus linguistics in a sense changes the priorities of philology as a science. “The object of study is speech, which cannot be reduced to linguistic abstraction, norms of literary language, judgments about correctness and incorrectness in language, based solely on the intuition of an educated researcher. The second important theoretical consequence of corpus research can be considered that the Saussure dichotomy is replaced by the idea of the primacy of speech activity with a smooth scale of generalizations from the cliché to the grammatical rule”. [4]

Computational linguistics, of course, provides the methodological basis for corpus research. Its importance has grown dramatically and dramatically with the growth and popularization of the Internet. Many researchers consider the WWW to be the largest and most powerful linguistic corpus, containing more than a billion documents. Two decades ago, a corpus containing a million words was considered large; nowadays corpora numbering more than one hundred million are being studied. [5]

Obviously, the use of the WWW as a language corpus is a new direction. The number of relevant approaches is quite limited, but tasks are successfully solved at different levels of the language system: syntax, semantics, as well as in the practical-translational aspect; in lexicography and translation. The presence of a variety of structurally and stylistically diverse texts in all standardized languages of the world on the Web, as well as artificial and fictional languages, provides researchers with almost limitless opportunities. It is also extremely important that the WWW contains parallel texts. There are new opportunities for diachronic study of language changes. And although outdated documents are often removed from sites, there are already successful results in studying semantic changes.

Experts in the field of theory and practice of translation believe that the WWW is the most useful tool for monitoring the specifics of the use of a word or phrase. Since the request can be limited by language or by country (because of URL), it becomes possible to obtain information about both the actual speech implementation and the frequency. Lexicography uses both vocabulary resources of the Internet and the detection of neologisms, their classification in different languages, the definition of valence, and so on.

From the very beginning, corpus linguistics has been closely associated with the teaching of a foreign language. It is known, for example, that 60% of English spoken language in the United States is accounted for by the 50 most frequent lexemes. Undoubtedly, this fact, obtained experimentally [6] should be taken into account when selecting the vocabulary at the appropriate stage of training. Educational dictionaries and textbooks of lexical and grammatical orientation are based on the results of corpus research. Unfortunately, such didactic materials have not yet been created for Ukrainian, Ukrainian-speaking audiences. A.K. Golovina’s textbook “Frequency course of accelerated learning of English in the field of radio electronics”, which was published in Leningrad in 1978 and was actively used in technical universities, is now obsolete, and has not found followers.

The creation of student test corpora is also promising, which would allow classifying errors, identifying typical ones and taking them into account in the teaching process. Such information is contained in some dictionaries (Collins Cobuilt Student’s Dictionary, Oxford Learner’s English Dictionary), but they usually do not take into account language interference. The creation of a textbook, which would take into account the mistakes of Ukrainian-speaking students when learning English, would be a step forward in the Ukrainian language didactics.

M. Wolf notes that the problem of the modern approach to the use of the WWW as a linguistic corpus is the lack of a specialized search engine. “We… have to live with the operators and options they offer. But these search engines are not tuned to the needs of linguists”. [5] The author even formulated requirements for an ideal search engine for linguistic corpus needs, including the possibility of limiting the search by syntactic, semantic, textual (stylistic), genre features, certain characteristics of the language unit (from a letter, a word to complex syntactic integer), etc.

Corpus linguistics has become increasingly popular in recent years, although it has been harshly criticized by the patriarch of modern scientific linguistics N. Chomsky [7]. Obviously, it is a full-fledged alternative to traditional philology.

Another area of interaction between the Internet and scientific linguistics is the study of Internet communication. It combines cognitive-pragmatic and linguistic-cultural aspects of study. Internet communication is interpreted today as a special environment for the actualization of language units, where the verbal component predominates. Experts talk about the formation of a specific language of the Internet, study its stylistic characteristics, the formation of genres (site, blog, chat).

Internet linguistics is sometimes understood as a "naive-scientific" study of Internet users, called to life by the text space of the WWW. This includes the creation of so-called fictional languages (conlangs), and "popular" translations, the creation of amateur dictionaries, attempts to organize terminology, and so on. The Internet, unlike other media, including television, involves complicity, is not a culture of consumption, and therefore formed the Internet language (Netspeak or e-talk), which is a multifaceted phenomenon for linguistic analysis.

Today, the Internet can be considered a symbol of our time, as its impact on the development of world civilization is unprecedented. It is a complex socio-technical system that is constantly working and changing, the popularity and importance of which is constantly growing. The study of the numerous consequences of this growth, as well as the internal space of the Network, requires an interdisciplinary approach, which is called Internet Studies. [8] Anthropocentrism of modern science brings to the fore the study of Web-space such humanities as sociology, psychology, computer science and, of course, linguistics.

References:

1. History of the Internet. https://en.wikipedia.org/wiki/History_of_the_Internet
2. Anthropocentrism. https://en.wikipedia.org/wiki/Anthropocentrism
3. Richard Nordquist. (2019) Definition and Examples of Corpus Linguistics. https://www.thoughtco.com/what-is-corpus-linguistics-1689936
4. M. V. Kopotev, A. Mustajoki. (2008) Modern Corpus Russian Studies. https://www.mv.helsinki.fi/home/kopotev/Kopotev_Mustajoki_2008.pdf
5. Martin Volk. (2002) Using the Web as Corpus for Linguistic Research. Publications of General Linguistic 3. University of Tartu.
6. Nation I. S. P. (1990) Teaching and Learning Vocabulary. New York.
7. Andor, József. (2004). The Master and His Performance: An Interview with Noam Chomsky. *Intercultural Pragmatics* – INTERCULT PRAGMAT. 1. 93-111. 10.1515/iprg.2004.009.
8. Constructed language. https://en.wikipedia.org/wiki/Constructed_language
9. Iryna Stepanova, Svitlana Nykyporets. (2021) Some functional-stylistic features of the modern scientific text. International scientific journal *Grail of Science*, (2-3), 338-340. https://doi.org/10.36074/grail-of-science.02.04.2021.069