ISSN 2227-5525. Мова і суспільство. 2017. Випуск 8. С. 5–11
Language and society. 2017. Issue 8. P. 5–11

УДК 81'33 (072)

COMMUNICATION IN SOCIETY: METHODOLOGICAL PRINCIPLES OF SOCIOLINGUISTIC RESEARCH

Lyubov Savelieva

Admiral Makarov National University of Shipbuilding
Heroes of Ukraine av., 9, NUOS, r. 418, 54000, Ukraine
semantema@gmail.com

The article is devoted to the study of the methodological basis of a new approach to empirical research in sociolinguistics which involves the use of computational methods in text analysis and can detect and handle sociolinguistic text data about patterns of language development, features of language functioning, specificity of language processes, important features of interaction and mutual influence of languages and other sociolinguistic factors. This approach is intended to promote the development of tools and methodologies of computational sociolinguistics that allows rapid disbursement in the information stream of linguistic facts and interpret them based on real, existing sociolinguistic text data.

Key words: computational sociolinguistics, sociolinguistic methods of text analysis, sociolinguistic automated text analysis, sociolinguistic text data.

The process of communication is a form of social action aimed at information exchange between people that determines the level and the type of organization of social groups. Successful communication in society is achieved if subjects (1) use the same channels of information dissemination (have the same access to information), (2) obey generally accepted laws and formal rules, conventions, customs, and traditions in communication. Today, this balance is disrupted in the Ukrainian society. The situation can be changed with the help of a complex, systematic, and qualitative multidimensional analysis of communication in a society that requires linguists to improve the methodology of sociolinguistic research.

With this in mind, we provide a comprehensive overview of modern tools for automated collection and processing of information that, consequently, will strengthen the methodological base of sociolinguistics. The analyzed means will allow to process large amounts of text within a minimum period and provide the most accurate results of research.

Modern Ukrainian sociolinguistics has a large number of works covering the methodological issues of research (B. Azhniuk, V. Demchenko, L. Masenko, H. Matsyuk, S. Sokolova, L. Stavytska, O. Tkachenko, V. Trub, K. Tyshchenko, H. Yavorska etc.), but the scientific literature still has not addressed sociolinguistic automation tasks. Therefore, the object of analysis, observation, and generalization of our study is the search for computational methods for text analysis, which can automate the procedures of identification and processing of *sociolinguistic text data.*

Note that *sociolinguistic text data* is an element of text that gives answers and explanations to the questions: (1) about patterns of language development under different socio-economic, ethnolinguistic and cultural conditions caused by the social nature of language; (2) about features of the communication process according to social roles and socio-psychological conditions of speech act implementations and their illocutionary force; (3) about the effects of interaction and mutual influence of languages under the condition of their co-existence in one society; (4) about problems of interference and borrowing of elements in language contact.

The task of automation of sociolinguistic research provides the deployment of the following areas: 1) algorithmization and software implementation of traditional methods of sociolinguistic research; 2) development of tools and methodologies for computational sociolinguistics to improve and speed up the analytical work with *sociolinguistic text data*.

1. Algorithmization and software implementation of traditional methods of sociolinguistic research

Automation of sociolinguistic research requires solving problems of algorithmic and software implementation of traditional methods of sociolinguistics. Algorithms have to be created for all traditional methods of linguistic research, namely the following methods: 1) collecting of material; 2) processing of the collected data; 3) assessment of the credibility of the results. A considerable number of these methods and procedures are in the current development of sociolinguistics already done by computer. Advantages and disadvantages of each of these methods depend on the type and number of analyzed texts and problems that linguists have been studying. However, there is no method that is perfectly suited for all kinds of linguistic data.

The *methods of sociolinguistic data collection* that can be fully or partially automated include 1) immediate (direct) observation; 2) survey; 3) interviews; 4) analysis of documentary sources; 5) testing. At the stage of data collection, demographic data are used: population, national, social, gender and age characteristics of the sample population, data about the development of education, culture, science, mass communication, productive activities, etc.

The material for the extraction of *sociolinguistic text data* can include magazine articles, headlines, keywords, different types of company reports, memos, letters, newspaper prescriptions, advertisements, public speeches, comments, blogs, chat rooms, scripts for video recordings, telephone talks, slogans, and memes.

The *methods of processing of the collected sociolinguistic data* that can be automated include procedures that linguists traditionally use in the analysis of source material. Data analysis involves the use of a statistical system, including the method along with scaling and correlation analysis carried out by computer.

2. The development of computer tools and methodologies of computational sociolinguistics

The second direction of automation in sociolinguistic research concerns the development of tools and methodologies of computational sociolinguistics. Today, this area is just beginning to develop, thus it opens up wide opportunities for future activities of sociolinguists.

Computational sociolinguistics covers a wide realm of computer instruments to create text annotation, description, and encoding of *sociolinguistic textual data* allowing to create an algorithm and, consequently, to automate this type of analysis.

The main task of computational sociolinguistics is the search for methods of automated sociolinguistic analysis of texts, which are practical and can provide reasonable conclusions and reliable results. At the current stage, we also need to ensure the creation of expert computer analysis systems, which offer choices in the selection of procedure sequences that perfectly meet the research requirements and allow checking the validity of the results.

As computational sociolinguistics is just beginning to develop its own tools, it is appropriate to use automated methods that have proved effective in areas such as a) communicative linguistics, b) corpus linguistics, c) computational linguistics, d) computational statistics, and e) computational sociology.

A) The use of automated methods of communicative linguistics in sociolinguistic research

To solve sociolinguistic tasks of algorithmization the methods that are used in **communicative linguistics** can be used, including discourse analysis, content analysis, transaction analysis, and intent analysis.

The main task of *discourse analysis* is to identify the social context and to study the relationship between language in communication and social, mental, psychological, and cultural processes [1: 26].

Content analysis, a set of techniques and methods to describe content and rules of communication, provides tools for assessing social studies, propaganda, communication psychology, the study of differences in culture, research in pathogenic and precedent texts and discourses, and is used for analysis of differences in speech and in communication behavior of mentally ill and healthy people [1: 27]. *Content analysis* is a multidisciplinary method that integrates the theories and methodologies of many humanities by the use of mathematical statistics and linguistics to study any given text.

Transactional analysis studies interpersonal relationships based on language and paralinguistic codes (gestures, facial expressions, posture, etc.); the analysis includes psychological games in which "ego-states" of speakers are found, and "scenarios" under which participants are unconsciously acting [1: 28].

The essence of *intent analysis* is to identify speech intents and goals of the speaker (public or hidden) [1: 29].

Automated discourse analysis, content analysis, transaction analysis, and intent analysis are based today only on identifying word forms and calculating the frequency of their use, which is not enough for sociolinguistic research. In addition, the content often cannot be analyzed without exclusion of the general context of situations described in the text and inclusion of additional linguistic information or the use of methods of natural language processing. The efficiency of content analysis of a given text by computer depends on the number of types of processed linguistic and metalinguistic information extracted from the analyzed text.

B) The use of automated methods of corpus linguistics in sociolinguistic research

To strengthen sociolinguistic research methodology, we can use the methods of corpus linguistics that allow the processing of large numbers of texts if special annotation (encoding) of sociolinguistic information is provided in the body of the texts. Today, there are many text corpora with linguistic markups, which can be used for sociolinguistic research. In addition, there are a number of electronic text archives containing text data from a large variety of sources that can be used for research purposes.

C) The use of sociolinguistic research methods in computational linguistics

A list of programs that are designed to perform computerized linguistic analysis of texts can be found in the "Catalog" section at the "Laboratory of Applied Linguistics of the National University of Shipbuilding" website [2]. Note that these directories contain direct links to resources available on the Internet containing: details for program developers; instructions how to work with them; conditions under which software is being distributed; the source files and libraries which allow you to create your own program with the necessary functions.

D) The use of sociolinguistic research methods in computational statistics

Modern computational statistics have developed powerful tools for automated analysis. In particular, there are *programs* for 1) *data classification:* BMDP New Syst, SigmaStat, Statistix, StatXact, Turbo Spring; 2) *for content analysis:* General Inquirer, AnnoTape, Cameo, Concordance, HyperResearch, JFreq, Lekta, Leximancer, Meca, MonoConc, Protan, QDA Miner, Qualrus, Salt, Stata, Statistica, Tabari (Keds), Textpack, TextQuest, TextSmart, Tropes, Words, WordSmith, WordStat, Baal, DA system, soprano, Evrysta, content analysis Pro (version 1.6), PLCA; 3) *for statistical data processing:* BM-STAT, DATA DESK, JMP, Multivariance, NCSS + Advanced Statistics, ODA, POWERSTAT, SAM- 86, SOLO, STADIA, StatView, Statit, STATlab, STATMOST, UNISTAT, WinSTAT, MESOSAUR, OLIMP, Rostand, SANI.

In addition, there are *special packages for data classification:* Cart, MVSP, Palmoda (LOREH), PolyAnalyst, Stars, Stat- Media, Quasar, Class-master; and *statistical expert systems:* Starex, Statistical Navigator, and STATЭKS.

All of these programs, special packages, and expert systems have advantages and disadvantages. Therefore, they can be valuable for certain purposes, but they may be unsuitable for other purposes. However, different applications require or combine different methods of analysis.

Functional modules of these programs, specialized packages, and expert systems provide the following operations: (1) data management; data conversion; (2) color graphics; (3) descriptive statistics; (4) t-criteria; (5) discriminant methods, variance, determination, cluster, covariance, component, correlation, regression and factor analysis of data; (6) methods of intelligence analysis, experiment planning and creation of hypotheses, hypothesis testing and analysis of stochastic processes (time series).

Automated methods of statistical analysis can reveal abnormal distributions and form a working hypothesis. In the subsystem of model construction, you can check the hypothesis; measure and compare the relationship of signs; identify and model the structure of dependence; analyze data that vary over time; determine the trajectory of objects; identify homogeneous groups; track the average time to go from one state to another; to simulate the process of evolution.

F) The use of sociolinguistic research methods in computational sociology

For an automatic sociological analysis of texts, the **methods of computer sociological analysis** of texts appear to be effective, such as the method of map analysis [3], the method of using semantic text grammars (method Franzosi) [4], the general inquirer method [5], the method of psychotherapy interview analysis [6], the method of contextual content analysis [7], the method of visualization of the content [8], the method of semantic and syntactic analysis of Nazarenko [9], the method of automated content analysis [10].

The method of map analysis [3] is aimed at analyzing text data with a set of categories and concepts including frequency and distribution of concepts in the text and the relationship between them. The method involves the analysis of content. This method is used to determine similarities or differences in the content and structure of a large number of texts dealing with certain topics and themes.

Using the method of map analysis, we get a grid of information in which concepts are nodes and "special values" reveal interactions between the nodes. The amount of information to be registered for each ratio is determined by the researcher.

The method of application of semantic text grammars *(Method Franzosi)* uses semantic grammar to identify categories and relations between them [4]. It includes also information about the type of text for more efficient coding. This method allows the user to perform an analysis of the content manually or by computer. The researcher is encrypting data manually and the computer is used only for calculations.

Unlike the analytic maps, the Franzosi method helps to find out relationships between concepts based on linguistic, rather than cognitive criteria. The method is applicable for corpora analysis with appropriate markings and texts of the same type.

The method of general inquirer [5] is implemented as a program that has online access. Functional modules of the program are providing an analysis of the content of the text and can be used to 1) identify concepts and media clusters that use them; 2) identify areas of influence; 3) identify contract or related publications, plagiarism; 4) dissemination of information; 5) search for experts; 6) search for necessary information; 7) identification of communities; 8) sampling study.

The program uses additional functional modules such as content analysis, discourse analysis, semantic nets, statistical analysis, cluster analysis, factor analysis, regression analysis, network analysis, dynamic network analysis, match analysis, multidimensional scaling, and analysis of variance.

The general inquirer method uses a dictionary of general research categories from four sources: the vocabulary of categories "Harvard IV", the semantic dictionary of Lasswell, the vocabulary of new categories and the vocabulary of category markers. For each category, there are lists

of words that express the meaning of these categories. It also uses dictionaries of prefixes and suffixes to identify indirect word forms present in the text. In this method, an algorithm compares each form of a word or phrase included in each dictionary and assigns them to the appropriate codes.

The method of psychotherapy interview analysis was specially designed for the analysis of communication between therapist and patient [6]. The origin of this method was the assumption that specific text data are a source of prior classification categories. The most important thing for this analysis is to study the relationship between the content and structure of verbal communication and the changes of the patient during psychotherapeutic treatment. In this method, all the words that remain after the removal of function words and lemmatizing and that have a frequency equal to or greater than ten are stored to screening. During the screening of correlations, random events occur that should not be seen, so they are removed following the criteria of significance appliance. The methodological advantage of this method is the attempt to save time searching for relevant concepts and their linguistic implementation. This method may also be useful in situations, which require finding common themes or categories in large amounts of text data.

The method of contextual content analysis [7] is designed to measure the social distance between positions (statuses) in the organization. It is based on the assumption that there are obvious stylistic differences between speakers, for example, between communication managers and employees or doctors and patients. This method has been implemented in the program *Minnesota* for contextual content analysis.

The method of content visualization [8] involves creating visual maps about the topics that dominate the text. The authors tried to overcome the problem of the method of reading and subjective interpretation and speed up the analysis process by automatically generating keywords occurring in the text. This means identifying important categories and their relations. This method can be applied in the analysis of media, where large amounts of text are compared and studied. Disadvantages of the method are: 1) a lack of "sensitivity" to the specific language because only graphemes are analyzed; 2) lemmatization is not made; 3) there are no ways to group semantically close words; 4) multiple meanings are not counted; 5) it is impossible to check whether some word groups (two or more) are used in the sense of one semantic unit, for example, an idiom. However, the method of content visualization can be useful for a specific analysis because it easily identifies the dominant themes and their relative importance.

The method of semantic and syntactic analysis of Nazarenko [9] is specifically designed for sociological research. This method relies on a quantitative analysis based on a syntactic and semantic analysis. The advantage of this method lies in the fact that statistical calculations are applied to the analysis of linguistic data, unlike standard statistical graphs. The authors used manual coding as well as automatic coding: links and semantic features were coded manually, while tags were set automatically. However, there is the well-known risk of inconsistent coding and big amounts of time and effort are required for manual labeling. The positive effect is that the syntactic information helps to identify patterns specific to each category of respondents, and allows to get results that cannot be easily removed from the statistics of graphical forms or segments. This method allows building a hierarchy of groups of words, which form hyponyms and synonyms, and it is a convenient way to represent conceptual information. The method of syntactic and semantic analysis can be used in a system of content analysis.

The method of automated content analysis [10] resulted from a cooperation between a university and a company research center on corpus linguistics. They created an automated content analyzer that combines specific linguistic and semantic markups. In this method, oral interviews are transcribed in text format, which allows marking up according to specific questions and features of the respondent (age, gender, social group, etc.). The next stage is performed by

drawing on the elements that indicate a certain behavior such as the pronouns "he" and "they." Automatic markup is done by using morphological tagging systems. In addition, each language unit receives semantic tags. Structurally ambiguous words are duplicated with separate entries for each part of speech. Polysemy is removed automatically using contextual rules. Manual post-editing is done for the rest of the tags and all uncoordinated points. The advantage is that this method uses linguistic information to identify relational links between different grammatical categories.

Thus, the article describes the methodological basis of the computational analysis of texts, which aims at identifying and processing sociolinguistic text data. The obtained results assert the feasibility of computerization of traditional research methods that will contribute to the development of tools and methodologies of computational sociolinguistics and allows rapid disbursement in the information flow of linguistic facts and interpret them based on real, existing sociolinguistic text data.

Conclusions

The described methods will help to improve the software for the implementation of methodologies for sociolinguistic analysis and to provide sophisticated modeling of text data according to a coding scheme. Of course, this software should improve the inclusion of linguistic information or use the dynamic call of external programs that will manipulate such information to supplement and correct it. In this way, it is possible to support the circular process of quantitative analysis, qualitative interpretation and adapt to different types of meta-information needed for deeper analysis. They impose technical requirements for complex systems and data modeling.

In the current stage isolated research on methodologies for analysis and on the creation of computer tools for text analysis are both economically and intellectually ineffective. Further perspectives for the automation of sociolinguistic lie in the cooperation with professionals of such disciplines, as sociolinguistics, computer science, sociology, and others.

1. *Бацевич Ф.С.* Основи комунікативної лінгвістики: підручник / Ф.С. Бацевич. Київ: Академія, 2004. – 344 р.
2. Catalogues of linguistic programs [electronic resource]. – Access: http://appling.in.ua/ pages/catalogs.php.
3. *Carley K.* Formalizing the social expert's knowledge / K. Carley // Sociological Methods and Research, 1988. – P. 165-232.
4. *Franzosi R.* Computer-assisted coding of textual data / R. Franzosi // Sociological Methods and Research, 1990. – P. 225-257.
5. General Inquirer [electronic resource]. – Access: http://www.wjh.harvard. edu /~inquirer/.
6. *Iker H. P., Harway N. I.* Words: A computer system for the analysis of content [electronic resource]. – Access: http://link.springer.com/article/ 10.3758/BF03200396/.
7. *McTavish D. G., Pirro E. B.* Pirro contextual content analysis. Quality and Quantity [electronic resource]. – Access: http://link.springer.com/article/ 10.1007/BF00139259.
8. *Miller M. M., Riechert B. P.* Identifying themes via Concept Mapping: a new method of content analysis. Presented to the Theory and Methodology Division, Association for Education in Journalism and Mass Communication Annual Meeting [electronic resource]. – Access: http://excellent.com.utk. edu/~mmmiller/ pestmaps.txt.
9. *Nazarenko A., Habert B., Reynaud C.* Open response surveys: from tagging to syntactic and semantic analysis. In Proceedings of JADT / A. Nazarenko, B. Habert, C. Reynaud // 3rd International Conference on Statistical Analysis of Textual Data. – Rome, Italy, 1995. – Vol. II. – P. 29-36.
10. *Wilson A., Rayson P.* The automatic content analysis of spoken discourse: a report on work in progress. Electronic reference: [electronic resource]. – Access: http://www.comp.lancs. ac.uk/ computing/research/ucrel/ papers/ war93.txt.

КОМУНІКАЦІЯ В СУСПІЛЬСТВІ: МЕТОДОЛОГІЧНІ ЗАСАДИ ДОСЛІДЖЕННЯ В СОЦІОЛІНГВІСТИЦІ

Любов Савельєва

Національний університет кораблебудування ім. адм. Макарова
пр. Героїв України, 9, НУК, к. 418, 54000, Україна
semantema@gmail.com

Статтю присвячено вивченню методологічних засад нового підходу до емпіричних досліджень у соціолінгвістиці, що передбачає застосування комп'ютерних методів аналізу тексту, за допомогою яких можна виявляти та опрацьовувати соціолінгвістичні текстові дані про закономірності мовного розвитку, особливості функціонування мов, специфіку процесів мовного спілкування, важливі риси взаємодії і взаємовпливу мов та інші соціолінгвістичні чинники.

Завдання комп'ютеризації соціолінгвістичних досліджень вбачаємо в розгортанні діяльності за такими напрямами: 1) алгоритмізація і програмна реалізація традиційних методів соціолінгвістичних досліджень; 2) вироблення інструментарію і методології комп'ютерної соціолінгвістики для покращення й прискорення аналітичної роботи з соціолінгвістичними текстовими даними.

Автоматизація соціолінгвістичних досліджень вимагає розв'язання завдань з алгоритмізації і програмної реалізації традиційних методів соціолінгвістики. Алгоритмізувати необхідно всі групи традиційних методів соціолінгвістичних досліджень, а саме: 1) збирання матеріалу; 2) опрацювання зібраних даних та оцінювання достовірності одержаних результатів. Значна кількість цих методів і процедур на сучасному етапі розвитку соціолінгвістики вже реалізується за допомогою комп'ютера.

Відповідно до другого напряму автоматизації соціолінгвістичних досліджень передбачається вироблення інструментарію та методології комп'ютерної соціолінгвістики. Комп'ютерною соціолінгвістикою називаємо напрям мовознавчих досліджень, який розв'язує теоретичні і прикладні завдання соціолінгвістики за допомогою комп'ютерних інструментів. Головним завданням комп'ютерної соціолінгвістики є пошук таких методів автоматизованого соціолінгвістичного аналізу тексту, які виявляються практичними та можуть забезпечити обґрунтовані висновки й достовірні результати.

Оскільки комп'ютерна соціолінгвістика лише починає виробляти власний інструментарій, то доцільним є застосування автоматизованих методів, що вже виявилися ефективними в таких галузях, як: а) комунікативна лінгвістика, б) корпусна лінгвістика, в) комп'ютерна лінгвістика, г) комп'ютерна статистика, д) комп'ютерна соціологія. Для розв'язання соціолінгвістичних завдань алгоритмізації піддають методи, які сьогодні становлять інструментарій комунікативної лінгвістики, зокрема: дискурс-аналіз, контент-аналіз, трансакційний аналіз та інтент-аналіз. Посиленню методології соціолінгвістичних досліджень також сприятиме використання методів корпусної лінгвістики, що забезпечить опрацювання великих масивів текстів.

Для автоматизованого соціологічного аналізу тексту ефективними можуть бути також методи комп'ютерного соціологічного аналізу тексту, такі як: метод аналізу карти, метод застосування семантичних текстових граматик (метод Франзосі), метод загального опитувальника, метод аналізу психотерапевтичних інтерв'ю, метод контекстного аналізу змісту, метод візуалізації змісту контенту, метод синтаксичного й семантичного аналізу Назаренка, метод автоматизованого аналізу змісту контенту.

Такий підхід покликаний сприяти виробленню інструментарію і методології комп'ютерної соціолінгвістики, що дозволить швидко виділяти в інформаційному потоці мовні факти й інтерпретувати їх на основі реальних, наявних у тексті соціолінгвістичних даних.

Ключові слова: комп'ютерна соціолінгвістика, методи соціолінгвістичного аналізу тексту, комп'ютеризований соціолінгвістичний аналіз тексту, соціолінгвістичні текстові дані.

Стаття надійшла до редколегії 10 січня 2017 року
Прийнята до друку 6 листопада 2017 року