B. Radić-Bojanić*
Faculty of Philosophy
Zorana Đinđića, Novi Sad (Serbia and Montenegro)
УДК 81'322

METHODOLOGY OF REGISTER ANALYSIS: COMPUTER-MEDIATED COMMUNICATION

The paper lists arguments in favour of compiling an annotated CMC corpus by illustrating how it could be used in the description of the methodology of register analysis, where such a corpus would be immensely useful. In addition, the paper lists some problems which will have to be dealt with during the compilation of such a corpus and states that they will have to be overcome if linguists want to conduct a thorough and detailed analysis.

1. Introduction

In the age of computers corpus linguistics has been dealing with the description and analysis of spoken and written texts in their electronic form. Surprisingly, communication in the electronic medium, which has made all this possible, has been largely neglected as regards corpus linguistics. Namely, despite the vast quantity of the spoken/transcribed and written texts which have found their place in electronic corpora, there have been very few attempts at including new modes of electronic communication (electronic chatting, e-mails, discussion group texts, etc.) into these corpora. This paper is meant to draw attention to this problem by showing how such corpora of computer-mediated communication (henceforth: CMC) could be used in linguistic research.

The current corpora greatly rely on the speech/writing dichotomy, which automatically excludes the new modes of electronic communication since they lie halfway between the two aforementioned poles. For instance, electronic chatting, which is formally written communication, displays a large number of oral characteristics. If CMC is to be understood fully and investigated in detail, the research will have to rely on an electronic corpus of texts, which would, in turn, be used in various types of analyses at different linguistic levels. Therefore, "it would seem well worth the effort to devise and gather well-sampled corpora typifying the new modes of communication, so as to understand their central and typical aspects and work towards the larger goal of unravelling the complex relationship between traditional forms of writing and speech, as well as between these traditional forms and the newer modes of communication, such as computer-mediated communication on the Web" [9: 92]. The process of compiling a CMC corpus needs to include two complimentary approaches: qualitative and quantitative. Whereas the qualitative aspect would deal with the circumstances in which communication takes place, the quantitative one would involve analyzing an electronic corpus with "existing part-of-speech software, in order to identify any unknown lexico-grammatical features, and [training] the software for future analysis of such modes of communication"[9: 92].

This paper deals with one form of CMC, namely electronic chatting. Through the description of the methodology of register analysis, the author hopes to show and prove the necessity of compiling a CMC corpus. After describing what is involved in register analysis, the paper goes on to consider the quantitative and qualitative aspects of electronic chatting and finishes with arguments which speak in favour of compiling CMC electronic corpora in various languages, not just of English.

2. Methodology of Register Analysis

"A communication situation that recurs regularly in a society (in terms of participants, setting, communicative functions, etc.) will tend over time to develop identifying markers of language structure and language use, different from the language of other communication situations" [6: 20]. Language structure is understood here as similar vocabulary, similar intonation features, similar syntax, etc. Register analysis, besides the typical linguistic characteristics of a register, must describe the extent of linguistic variability, e.g. the range of syntactic structures or intonations which occur in a communication situation. Studies of register have three major components [2: 10]:

- description of the situation in which the register is used
- description of the linguistic features of the register

- analysis of the functional or conventional associations between the situational and linguistic features.

These relationships are bidirectional: situational characteristics influence the choice of linguistic forms, while the choice of linguistic features helps to create the situation [2: 10].

Register analysis is conducted in several stages, which are described in detail in Biber [1: 64]. As is stated there, the preliminary analysis considers previous research in order to identify potentially relevant linguistic features. After that, a corpus of texts, which is to be used in the analysis, is compiled. Alongside with the compilation of a corpus, note must be made of all relevant characteristics of the situation in which the texts were produced. At the end of the preliminary phase, various linguistic features of the text are singled out and counted. With already annotated corpora this part of analysis is done fairly quickly, but the problem arises with the texts which are not annotated. This is the case with CMC, which is a paradox. On the one hand, all texts of CMC are readily available because of the very nature of the communication process – it takes place exclusively via computers. On the other hand, they are not annotated, which present a great problem in any kind of thorough linguistic research.

The very annotation of an electronic chatting corpus will encounter certain problems, connected with CMC-specific features, such as abbreviations, emoticons, typos, respellings, etc., illustrated in the text below:

<*Shania> I just d/led at 405 k/s rofl !!!!!!!!!!!!!!* – abbreviations
<*Siren> OMG!* – abbreviation
<*Rom|Breakfast> wb Shania sorry for being a smart idiot b4* – abbreviation, respelling
<*Rom|Breakfast> omg* – abbreviation
<*Sphinx30> woohoo*
<*Siren> cool*
<*ICU2> Oh my I missed you so much.....NOT :P* – emoticon
<*Waiter|v|> :)* – emoticon
<*Shania> Sphinx30 i wanna send u soemthing* – respelling, abbreviation, typo
<*Shania> ti see* – typo
<*Shania> and then u send me soemthing* – abbreviation, typo
<*Sphinx30> hmm ok*
<*Shania> 22 both ways for us Sphinx30*
<*Sphinx30> looks like it*
<*Sphinx30> i wish it would be faster*
<*Shania> try this Sphinx30*
<*Sphinx30> but its fast*
<*Shania> type /pdcc 20000*
<*Shania> why isnt it fast to Sphinx30 yet it is elsewhere*
<*Shania> Sphinx30 is egtting it at 22 k/s* – typo
<*Sphinx30> you did shoot it by mirc elsewhere Shania*
<*RED^EviL> HIIIIIIIIIIIIIIIIIIIIIIIIIIII* – emphasis
<*HomeAlone> me to.....*
<*HomeAlone> :))))))))))))* – emoticon

Besides that, a CMC corpus should contain information about punctuation or the lack of it, since that also contributes to determining the features of this register. As can be seen in the text above, punctuation is not used in the traditional way. Rather, it becomes part of the new CMC conventions, where punctuation marks function as a means of emphasis or building blocks for emoticons. In addition, capital letters have lost their traditional roles (capitalizing the personal pronoun "I", personal names, sentence beginnings, etc.) and they are also used in new ways, again mainly to achieve emphasis.

In the next phase of register analysis the linguistic features are grouped according to their frequency of occurrence in the texts and then are interpreted by assessing their communicative functions. In other words, a frequent occurrence of the passive voice points to a more formal style, usually found in scientific texts. Moreover, a frequent occurrence of personal and demonstrative pronouns speaks of the importance of the connection between the text and the extralinguistic context, especially in interpreting the reference of the pronouns.

The last phase, which is particularly important in the analysis of different kinds of texts and registers, compares statistical data for linguistic features and interprets their link with the texts and situations in which they are found more or less frequently.

In order to arrive at a description of the system of registers in a language, as well as to correctly explain the occurrence of linguistic features in some types of texts, a comprehensive register analysis must be based on three equally important factors: the description of a situation, representative samples of texts and a detailed description of linguistic features occurring in a

register. Biber [1: 70] states that texts differ on the basis of topics, purposes, rhetoric structures and styles, as well as situational parameters, such as the relationship among the participants in the communication process, the attitude of participants towards the extralinguistic content and the attitude of participants towards the text itself. Therefore, one of the mistakes that can occur in corpus compilation is not making notes of the participants in the communication process and the circumstances of text production [1: 70]. This causes mistakes in register analysis, since some linguistic features, without the information on extralinguistic factors, can be interpreted in a wrong way. That is why it is necessary for researchers to include information about the extralinguistic elements of the communication situation while compiling a corpus. To conclude, every thorough register analysis must be based on two very important methodologies: statistics, which testifies of the frequency of occurrence of linguistic features, and the ethnography of communication, which interprets the results of the investigation taking into consideration the extralinguistic and situational factors.

The quantitative analysis, conducted by way of statistics, interprets the differences among registers on the basis of differences in the distribution of linguistic features, which means that a relative frequency of several features identifies a register. The statistical processing of the data has a few phases [1: 75–78]. First of all, a choice is made of linguistic features which should receive special attention in a particular corpus or a particular kind of research. The researcher relies on the previous studies while making the choice of the features, which is part of the preliminary phase of the register analysis. In order to get as many details in the analysis as possible, a large number of potentially important features should be included. The counting is based on a text of 1000 words or on the texts of the same length so that the comparison is correct. In case the comparison is done on the texts of different lengths, the frequency data will not be correct. Therefore, according to Biber [1: 76], the formula for calculating the frequency of a certain linguistic feature is as follows:

number of occurrences of one feature/number of words in a text · 1000 =
number of occurrences of the feature in a text which is 1000 words long,
i.e. the frequency in 1000 words

Besides the quantitative analysis, there is a need for a non-quantitative one. While the former supplies an empirical background, the latter is necessary in order to interpret the results correctly. In other words, single linguistic features are neither enough to establish the differences among registers nor to describe one register in detail. On the contrary, the co-occurrences of several features and their interrelations are the core of the register description. Biber [1: 70] and Biber & Finegan [3: 7] stress the importance of the knowledge of the context in which the text was produced, since registers correspond to a series of activities characteristic of members of a community. That is why the ethnography of communication is a good starting point, since it studies the immediate use of language in a context. A social community in this case is the context, so its communicative activities are studied in the whole [7: 15–16]. The ethnographic practice is based on observing the communicators, on stressing the importance of the specificity of social life and points of view of other participants in the communication process [7: 23], which means that the researcher must spend a longer period in the field, must become part of the environment he/she is investigating, must become acquainted with the connections and activities and finally start understanding the environment and the participants [8: 4–5]. The researcher, on the one hand, must be close enough to the culture he/she studies in order to understand how it functions, while, on the other, must be capable of distancing him/herself enough in order to be able to write and report about it. Considering the involvement of the researcher in the situation, the problem of objectivity arises immediately and the only way to present the results in an objective way is to rely on the quantitative analysis, which in turn substantiates or refutes ethnographic conclusions.

The ethnography of the Internet is not necessarily connected with a physical journey, as was the case with earlier ethnographic research, where valid results could not be achieved with a certain amount of field work. On the contrary, ethnographic Internet studies focus on an experiential, not physical journey [8: 45], since one travels on the Internet by looking, reading, imagining and experiencing. The very lack of physical presence in the situation which is investigated introduces the element of unreliability of the source of information and the data which the researcher can gather and later analyze. This unreliability occurs due to the fact that it is impossible to check the information, that the real identities of communicators are not known to researchers and that the membership of a group or community is not constant.

3. Aims of Register Analysis

The aim of register analysis is to investigate the link between the linguistic expression and the social situation, with a view towards explaining that link [3: 3–12]. In the same manner, the analysis of the register of electronic chatting aims at exploring and explaining the following:

- how the characteristics of the situation in which CMC takes place change the language and to what extent;
- what linguistic and extralinguistic features make this register different from other registers;
- is there a connection between extralinguistic features (e.g. lack of face to face contact, keyboard and computer screen as the only ways of producing and receiving the information) and linguistic features (e.g. typos, innovativeness and brevity of expression).

Finally, the aim of register analysis is to identify and interpret the generalizations concerning register variations, as well as to systematically list linguistic and extralinguistic features of the register [2: 29–30]. A point of importance here is that the linguistic features must be interpreted not only in relation to the extralinguistic features of text production, but also in relation to other linguistic features, because the isolated analysis of a linguistic feature may lead to wrong conclusions. Thus, the use of personal pronouns of the first and the second persons singular, direct questions and imperatives indicate interactivity, without disclosing if they are found in a written or a spoken text, while abbreviated forms and self-corrections point to a spontaneous spoken discourse.

4. Conclusion

After presenting a number of reasons for compiling an electronic annotated CMC corpus, it must be said that this is a task that should be done in as many languages as possible. This is because the interlinguistic comparison and contrasting reveals explicit similarities and differences between two or more languages and may draw attention to phenomena that would otherwise be missed. Also, that kind of analysis would reveal any linguistic and cultural differences in one register across two or more languages, thus contributing to a better understanding of other languages and cultures.

References

1. Biber, D. (1988). Variation across Speech and Writing. Cambridge University Press. Cambridge.
2. Biber, D. (1995). Dimensions of Register Variation. A Cross-Linguistic Comparison. Cambridge University Press. Cambridge.
3. Biber, D. & Finegan, E. (1994). "Situating Register in Sociolinguistics". In: Biber, D. & Finegan, E. ed. Sociolingustic Perspectives on Register. Oxford University Press. Oxford. (3-12).
4. Crystal, D. (2001). Language and the Internet. CUP. Cambridge.
5. Danet, B.(2001). Cyberpl@y. Berg. Oxford, New York.
6. Ferguson, Ch. E. (1994). "Dialect, Register and Genre: Working Assumptions about Conventionalization". In: Biber, D. & Finegan, E. ed. Sociolingustic Perspectives on Register. Oxford University Press. Oxford. (15–30).
7. Hajmz, D. (1980). Etnografija komunikacije. [Foundations in Sociolinguistics. An Ethnographic Approach.] Beogradski izdavačko-grafički zavod, XX vek. Beograd.
8. Hine, C. (2000). Virtual Ethnography. SAGE Publications. London, Thousand Oaks, New Delhi.
9. Ooi, Vincent B. Y. (2000). "Aspects of Computer-Mediated Communication for Research in Corpus Linguistics". In: Peters, P., Collins, P. & Smith, A. ed. Language and Computers, New Frontiers of Corpus Research. Papers from the Twenty First International Conference on English Language Research on Computerized Corpora Sydney 2000. (91–104).

Р. Ющенко, В. Гудзенко*
Институт кибернетики им. В. М. Глушкова НАН Украины (Киев)
УДК 81’322.33

ПРИМЕНЕНИЕ ПАРАЛЛЕЛЬНЫХ КОМПЬЮТЕРОВ ДЛЯ АВТОМАТИЗАЦИИ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ

Морфологическое аннотирование Национального корпуса предполагает значительные усилия, учитывая объем текстов, который составляет сотни миллионов слов. Большую часть рутины, однако, можно избежать, подключив к работе современные компьютерные технологии. Задача морфологической разметки поддается частичной формализации, а следовательно, процесс тегирования можно автоматизировать. В рамках этих работ Институтом украинского языка были