Slovo. Journal of Slavic Languages, Literatures and Cultures
No. 65, 2025

Sarajevo Corpus of SMS Messages in Bosnian

Azra Hodžić-Kadić
Sprachenzentrum der Universität Wien / Institut für Slawistik und Kaukasusstudien, Friedrich Schiller-Universität Jena

azra.hodžić-kadić@univie.ac.at

Azra Ahmetspahić-Peljto
Institut für Slawistik und Hungarologie, Humboldt Universität zu Berlin

azra.ahmetspahic-peljto@hu-berlin.de

Today, we live in a world of constant correspondence using our mobile phones: we regularly send messages through numerous applications for which it is enough to have access to the Internet and then nothing limits us in typing messages. However, do we remember a time when it wasn't like that? Not so long ago, thirty years ago, on December 3, 1992, the first SMS message was sent. From the first message to the appearance of applications such as WhatsApp, Viber, Facebook, Instagram – civilization communicated via SMS and through this communication developed a language that had its own specifics.

The language of SMS messages is defined primarily by text limitations: one message could have a maximum of 160 characters with spaces. Therefore, users were forced to make their language as economical as possible, that is, to convey as much meaning as possible with as few language symbols as possible, regardless of the possible formality imperatives. The conventions of using this language are not isolated: they have left a strong impact on our entire correspondence, so even our current communication through a large number of applications is strongly influenced by former SMS messages. At the same time, the language of SMS messages itself developed together with the technology that enabled later certain corrections and suggestions of words.

This is why this language is interesting both to linguists and to researchers of other non-linguistic disciplines, such as sociology, psychology, cultural studies, etc. However, despite the great potential that SMS messages offer, SMS message corpora are rare in the world, primarily due to the difficult process of collecting samples that will be diverse enough to cover different aspects of interest to researchers. Most of these corps are privately owned and are not available to the public today. Among the active corpora available to the scientific community, the following can be distinguished: The Swiss SMS Corpus, 88MilSMS: French text-message corpus and sms4science.

Recognizing the potential that such a corpus would offer, a group of authors gathered in Sarajevo and decided to use perhaps the last chance to "catch" SMS messages in the

Bosnian language and create a corpus before they, together with "non-smart" phones, are forever gone.

The author's team consists of six linguists: Halid Bulić (project leader), Elma Durmisević, Azra Hodzic-Cavkic, Enisa Bajraktarević, Azra Ahmetspahić-Peljto and Belmin Sabic. The idea was formally developed as a project of the Center for Bosnian, Croatian and Serbian Languages of the Faculty of Philosophy of the University of Sarajevo, under the name *Sarajevo Corpus of SMS Messages in Bosnian*, not because the informants or authors are exclusively residents of Sarajevo, but simply because the idea of the project was born and developed in the capital of Bosnia and Herzegovina. Work on collecting SMS messages for the corpus began in January 2021, when the authors set themselves the imperative to collect 10,000 SMS messages. In addition to the fact that corpora of SMS messages are very rare and, from today's point of view, particularly precious, it is important to point out that the Bosnian language does not have many electronic corpora in general. Furthermore, they were all either developed outside the borders of Bosnia and Herzegovina, or are of the open or specialized type. Therefore, the *Sarajevo corpus of SMS messages in Bosnian* represents a very significant contribution to the development of corpus linguistics within the framework of Bosnian studies.

In February 2023, they succeeded in this and published the collected material in PDF format on the website of the Faculty of Philosophy of the University of Sarajevo[108]. Although the PDF format was not an ideal option, it still seemed a pitty to keep the material from the public. In this regard, project leader, Prof. Dr. Halid Bulić, wrote in the *Preface* about this electronic edition:

> The Sarajevo Corpus of SMS Messages in Bosnian was originally published by the University of Sarajevo – Faculty of Philosophy as an electronic book. The second phase of the work involved compiling the SMS messages into a corpus and linguistic annotation, which was done using the CLASSLA package (https://github.com/clarinsi/classla), version 2.1, with language = Serbian and type = nonstandard for tokenization, lemmatization and morpho-syntactic tagging (both MULTEXT-East and Universal Dependencies). As opposed to the previous version, this version corrects a number of mistakes in the metadata.

After the publication of the material in PDF format, work on the idea continued with the aim of getting the corpus into an adequate electronic form that will allow users different types of analysis, both linguistic and non-linguistic. Therefore, in July 2024, in cooperation with Dr. Philipp Wasserscheidt, an expert in corpus linguistics from the Humboldt University in Berlin, found the corpus on the repository *CLARIN.SI*[109]. This corpus consists of 10,000 SMS messages, and can be counted among smaller corpora,

[108] The PDF can be found at the link: https://www.ff.unsa.ba/index.php/bs/projekti-centra-za-b-h-s-jezik/18335-sarajevski-korpus-sms-poruka-na - in the Bosnian language.

[109] The corpus can be found at the link: https://www.clarin.si/repository/xmlui/handle/11356/1913#:~:text=The%20Sarajevo%20Corpus%20of%20SMS%20Messages%20in.

but the variety of its samples certainly offers a large number of possibilities for research. Here it is stated that the corpus contains: *10000 texts, 15330 sentences, 105902 words, 128492 tokens* and the description further says:

> This corpus is specialized, static (i.e., no future growth is planned), diachronic and covers the period from 2002 to 2022. The SMS messages included in this corpus were obtained from voluntary donors (informants). Both senders and recipients of the messages included in the corpus are Bosnian speakers, exhibiting diversity in terms of age, education and occupation, place of origin and countries of long-term residence. The Sarajevo Corpus of SMS Messages in Bosnian was originally published by the University of Sarajevo – Faculty of Philosophy as an electronic book. The second phase of the work involved compiling the SMS messages into a corpus and linguistic annotation, which was done using the CLASSLA package (https://github.com/clarinsi/classla), version 2.1, with language = Serbian and type = nonstandard for tokenization, lemmatization and morpho-syntactic tagging (both MULTEXT-East and Universal Dependencies). As opposed to the previous version, this version corrects a number of mistakes in the metadata.

Each SMS in the corpus offers information about the age and occupation of both the recipient and sender of the message, place of residence, level of education, gender, number of words in the message and number of characters. In addition, the informants were able to state notes that they considered important, for example that a particular message is a continuation or response to a previous message or that the recipient and the sender are related. Therefore, within the framework of this correspondence, users can analyze linguistic and non-linguistic aspects of entire love correspondences, discussions, conflicts or, on the other hand, explanations of why the informant is late for the arranged coffee, etc. The anonymity of the informant is guaranteed because all names, surnames and nicknames in the text of the message are anonymized. Within the 10,000 SMS, the informants are very diverse: they come from different age groups (from students to pensioners), different educational profiles, and even different countries. Although the informants are mostly from Bosnia and Herzegovina, a certain number also come from Croatia, Montenegro, Serbia, Slovenia, Sweden, Austria, Germany and the United States of America. The condition for specifying another country was that the informant spent more than five years there, so this corpus can also be used for the analysis of bilingualism or language in a diaspora. The messages contained in the corpus were sent in the period from 2002 to 2022, so this corpus also has a diachronic nature because it can be used to analyze how SMS correspondence has changed over time.

The material for the corpus was collected by having informants fill in a table for each SMS they shared. In the text of the message, the name is anonymized, but there were no other interventions. The emoji has been left. Metadata includes: native language (which is the same for all informants - Bosnian), gender, age, country of long-term residence, level of education, occupation, and month and year of sending the message. Taking into account the above data, the search on the Clarin.si repository offers numerous possibilities. We can examine, for example, phenomena related to gender or age, for example the relationship between the number of words and the level of education, the choice of the most common initial wording among female and male informants, etc.

SMS, which served as the basic form of correspondence for about thirty years, is certainly a valuable phenomenon that science should record, which is why this corpus represents a very significant contribution both to Bosnian studies and beyond. Bearing in mind that the non-standard language is still fighting for its place in the study of the languages of the Central South Slavic diasystem and that this is a corpus of a non-standard language – it is important to highlight its regional and general Slavic contribution. And finally, given the variety of metadata, it can be confirmed with certainty that the *Sarajevo corpus of SMS in Bosnian* can be used by linguists in their studies of sociolinguistics, stylistics, pragmatics, dialectology or gender studies, but certainly also by other scientists within sociology, psychology, pedagogy and other sciences for which this data is relevant.