https://doi.org/10.18485/vlings_vele.2025.ch11
323.15:811(497.11)
81`26/27(497.11)

Dušan Vlajić[*], **Valentina Sokolovska**[**]

SAMPLING WITHIN THE PROJECT "VULNERABLE LANGUAGES AND LINGUISTIC VARIETIES IN SERBIA"

Abstract: The main objective of this chapter is to point out the characteristics of samples and the methods of their formation, both in the pilot and the main research, conducted within the project *Vulnerable Languages and Linguistic Varieties in Serbia* (*VLingS*). The project aimed to create a sociolinguistic questionnaire to assess the vulnerability of languages spoken by certain linguistic communities. These languages had previously been identified as endangered. Specifically, the questionnaire was administered to members of Roma, Vlach, Bayash, Rusyn, Banat Bulgarian, Aromanian, and Ladino communities. One of the major challenges was defining the population boundary for vulnerable languages, especially those with a small number of speakers and not visible in Official Census Publications. Therefore, the sample was formed based on various sources: the results of the 2011 census conducted in the Republic of Serbia, international databases (e.g. UNESCO, Catalogue of Endangered Languages, Ethnologue), and mostly based on the researchers' information harvested from own previous fieldwork and domestic literature. The sample represents a combination of three non-probabilistic sampling methods – purposive, quota and snowball sampling. The purposiveness of sampling is reflected in the selection of respondents,

[*] University of Niš
Faculty of Philosophy
dusan.vlajic@filfak.ni.ac.rs
https://orcid.org/0000-0003-0084-365X

[**] University of Novi Sad
Faculty of Philosophy
valentina@ff.uns.ac.rs
https://orcid.org/0000-0002-4106-6212

while the quota refers to determining the number of respondents within each specific linguistic community. The specific number of respondents overall (at the level of the entire sample) is defined using specialised software (GPower), taking into account the statistical analysis that will be subsequently conducted. For the selection of specific participants, the "chain referral" aka "snowball sampling" was chosen. The chapter also discusses the implications of conducting research on groups with a different number of respondents on statistical processing and analysis.

Keywords: sampling, sample size, data sources, vulnerable languages

1. Introduction

The primary objective of this chapter is to explain the methodology underlying both the pilot study and the main research on vulnerable languages and linguistic varieties. Both studies were conducted in Serbia in 2022 and 2023 within the *VLingS* (*Vulnerable Languages and Linguistic Varieties in Serbia*) project. The chapter focuses on several key methodological aspects. First, it outlines how the target population of the research was defined or, more specifically, the criteria used to decide which languages/varieties fall into the endangered category. Next, we describe the sample selection process. This section details the type of sampling method employed and the reasoning behind the choice to base the sample on non-probabilistic sampling. The following section covers the individual types of non-probabilistic samples that our sample relies on. In other words, the sample used in this study can be described as a combination of several types of non-probabilistic samples. Specifically, one sampling method was applied at the initial stage after we defined the target population, another once the number of individuals that needed to be surveyed within a specific language community was determined, and a third when it became necessary to reach certain individuals within a given community. Finally, we include the field researchers' observations and examples of guidance received from participants on selecting specific respondents.

The final section of the chapter provides a detailed description of the approach we relied on to determine the total number of participants. This topic was of particular importance to our project and received considerable attention during the preparation of the research. The primary reason is that the project was designed to include both qualitative and quantitative

analysis of the collected data. In other words, alongside interviewing participants, permanently recording and archiving their voices as they speak in endangered languages, and asking open-ended questions subjected to qualitative analysis, another aim of the project from the outset was to create a questionnaire. This questionnaire comprises questions related to various aspects of endangered languages (general data about linguonyms and language use, data on language acquisition and intergenerational language transmission, domains of language usage, etc.; more on this in Mirić, Sokolovska & Sorescu-Marinković 2024). Each of these aspects was operationalised through multiple individual questions, with each question coded as a separate variable when entering data into the SPSS database. The potential interrelationships among variables were tested using various statistical tests. For these tests to be meaningful—that is, to possess sufficient statistical power to yield statistically significant results even with small effect sizes—the minimum number of participants in the entire sample needed to be precisely determined. For this purpose, we used specialised software (GPower 3.1). Given the diverse types of statistical tests planned for data processing, several different methods for estimating the minimum number of participants had to be combined. As mentioned earlier, the entire procedure will be described in greater detail in the final section of the chapter, along with illustrations and concrete examples.

2. Basic information about the VLingS project and endangered languages

Minority languages have unique and complex statuses that must be considered during their study (Mirić, Sokolovska & Sorescu-Marinković 2024). One aspect of the complexity of minority languages concerns their classification into categories of endangered or vulnerable languages (for a detailed discussion on classification issues, see Sorescu-Marinković, Mirić & Ćirković 2020), focusing on issues such as the number of speakers (and, in some cases, even the enumeration of speakers) or their endangerment level. The VLingS project, launched in January 2022 with support from the Science Fund of the Republic of Serbia, has addressed these and other related issues.

Since one of the project's objectives was to highlight problematic classifications and correct data in international databases regarding

endangered and vulnerable languages in the Republic of Serbia, the study focused on:

- Aromanian,
- Banat Bulgarian,
- Vojvodina Rusyn (Ruthenian),
- Judezmo (Ladino),
- Romani (Vlax and Balkan varieties),
- Megleno-Romanian,
- Vlach, and
- Bayash Romanian.

The specificity of the VLingS project's theme and the sample involved in this research necessitated an appropriate approach to methodology and data collection. The uniqueness of the sample is reflected in several aspects. First, the project's target population is geographically dispersed, with members of different language communities living in various parts of the country. Certain groups (e.g. Roma) live across all regions of the Republic of Serbia. Consequently, the geographic distribution of locations where we conducted research on the Romani language had to align with the overall distribution of the Romani population in the country (for more details on the locations where the research was conducted, see Mirić, Sokolovska & Sorescu-Marinković 2024). In contrast, smaller language groups are concentrated in certain areas of a single region (Vojvodina), necessitating a different approach when reaching out to respondents.

The very fact that the VLingS project includes speakers of languages and linguistic varieties previously assessed as endangered makes members of these groups atypical representatives of the Republic of Serbia's population. Another important aspect is that these language communities have varying numbers of members. According to official census data conducted by the Statistical Office of the Republic of Serbia in 2011 (Đurić et al. 2014), the number of speakers ranges from a few dozen Aromanian speakers to over 100,000 whose mother tongue is Romani (Table 1). For some of these language communities, the official census does not provide speaker numbers (e.g. Megleno-Romanian or Bayash Romanian). On the other hand, according to unofficial information held by the project team members, obtained through fieldwork and a review of specialised literature, these numbers are markedly different. More detailed data are presented in Table 1.

Language/ variety	2011 Census (number of speakers)	2022 Census (number of speakers)	Unofficial Estimate	VLingS Pilot (target)	VLingS Pilot (realised)	VLingS Main (target)
Romani	100,668	79,687	significantly more	120	62	360
Vlach	43,095	23,216	150,000–300,000	60	44	180
Rusyn	11,340	8,725	/	30	30	90
Banat Bulgarian	601	/	between 1,000 and 4,000	10	8	30
Aromanian	243 (number of Aromanians)	327[1]	/	10	5	30
Ladino	/	/	less than 10	at least 1	1	10
Megleno-Romanian	/	/	several/ unknown	at least 1	1	10
Bayash Romanian	/	/	several tens of thousands	30	7	90

Table 1: Estimated number of speakers of linguistic varieties and target number of respondents for each language group.

This selection of languages involved unique methodological challenges in certain research phases (for more on methodological inquiries during data collection, see Radić-Bojanić & Sokolovska 2024).

The most commonly used instruments for assessing language endangerment are: the Graded Intergenerational Disruption Scale (GIDS) (Fishman 1991), the Extended Graded Intergenerational Disruption Scale (EGIDS) (Lewis & Simons 2010), the Language Vitality Index (Brenzinger et al. 2003), Degrees of Language Endangerment (Krauss 2008), and the Language Endangerment Index (LEI) (Lee & van Way 2016). In previous studies addressing endangered languages, these instruments have been applied within specific linguistic communities, whose number of speakers

[1] The 2022 Census reports the total number of Aromanians, but the number of speakers is unknown.

varies from one community to another. Moreover, the questionnaires are not administered through interviews with individual speakers in the field but are instead completed based on existing data about the community in question (e.g. number of speakers, presence of the language in education or media, domains of use—formal or informal, etc.).

Our project is unique in that it:

1) involves several distinct linguistic communities within the territory of a single country, making the sampling approach incomparable to prior studies on endangered languages;
2) was conducted within specific communities through interviews with individual speakers (see Mirić et al., this volume).

On the other hand, there are instruments that engage individual speakers directly, examining degrees of bilingualism, as well as the linguistic profiles of bilingual or multilingual speakers, including frequency and domains of language use. These instruments are often validated through testing specific speakers, and the number of participants in such validations also varies. Examples include LEAP-Q (Marian, Blumenfeld & Kaushanskaya 2007), the Language History Questionnaire (LHQ) (Li, Sepanski & Zhao 2006), and the Language and Social Background Questionnaire (LSBQ) (Anderson et al. 2018). However, these studies do not provide detailed information about participant recruitment. For instance, Marian, Blumenfeld and Kaushanskaya (2007) merely note that participants were recruited from a particular university campus and geographic region, whereas Anderson et al. (2018) specify that their study included participants from studies previously conducted by one of the co-authors. This is yet another reason why the sampling methodology employed in the VLingS project cannot be directly compared to that of prior, similar studies.

Since this chapter focuses on explaining the sample design for the survey, we will henceforth direct attention specifically to the sample's characteristics and design.

3. Basic information about probability and non-probability sampling

Considering all previously mentioned specifics of the project and the population it covers, a non-probabilistic sampling method emerged as the logical and only feasible method. This sampling method is characterised by

selecting participants based on the researchers' judgments and the specific needs and objectives dictated by certain studies (Milas 2009). The primary characteristic of such samples is that researchers cannot determine whether there is any probability that an individual will be included in the sample, and, if there is, what that probability is. In other words, it cannot be claimed that such samples are unbiased or representative of the population of interest. This is the main distinction from probabilistic samples, where the probability of selecting an individual for the sample is known to the researcher. While this probability does not need to be equal, it is nonetheless known and non-zero. In practice, this means that researchers have a sampling frame – an actual list of individuals from which the sample will be drawn. In the next step, researchers follow specific rules and procedures unique to certain types of probabilistic samples and select members of the population who will form the final research sample. Due to these factors, probability samples allow for the application of statistical techniques to estimate variable values and relationships at the population level.

The literature identifies simplicity, cost-effectiveness, lower expenses, and shorter research durations as advantages of non-probabilistic samples (Milas 2009). In certain situations, this is also the only feasible way to form a sample. This study falls precisely into the category in which non-probabilistic sampling is not only the only viable approach but also the only justified one. The reason is the previously mentioned 'atypicality' of the respondents included in the research, primarily because, apart from Serbian, they also use or are familiar with at least one other language and are, as such, the target group of this study. On the other hand, the vast majority of Serbia's population listed Serbian as their mother tongue in the 2011 census. However, this portion of the population was not the focus here. These two factors made reliance on probabilistic samples irrelevant for this project. In other words, the VLingS project was not intended to examine the linguistic habits and characteristics of the entire Serbian-speaking population. For such a purpose, a probabilistic sample, representative of the whole population, would be suitable. On the other hand, for a study defined in this way, a non-probability sampling method is far more appropriate. Furthermore, one of the goals of the VLingS project was to develop a sociolinguistic questionnaire to assess the degree of vulnerability of the languages used by certain communities. Thus, the emphasis is not on generalising findings to the entire population but on assessing specific linguistic parameters within

particular subpopulations. This is yet another reason why non-probabilistic samples were chosen.

The relevant scholarly literature distinguishes several types of non-probabilistic samples: convenience, quota, purposive, and snowball sampling (Milas 2009). The VLingS project used a combination of several samples from this group, which some authors refer to as mixed sampling (Fajgelj 2010). Specifically, our sample was obtained by combining purposive, quota, and snowball sampling.

4. Purposive sampling

In purposive sampling, respondents are selected based on prior knowledge about the population or defined research objectives (Milas 2009). This means that the composition of the sample is tailored to the researcher's assessment of possessing a specific critical characteristic or meeting a particular research purpose. The critical characteristic within the VLingS project was the assessment of the vulnerability of linguistic varieties. In other words, the primary criterion for determining whether a respondent would become part of the sample was the assessment of whether they belong to a community that uses a language previously classified as vulnerable (more on this in Mirić, Sokolovska & Sorescu-Marinković 2024).

There are several ways for researchers to select respondents when assembling a purposive sample. One method is the selection of modal members (Fajgelj 2010). Modal members are typical representatives of a particular group, i.e. individuals who possess a certain characteristic considered to best represent the phenomenon of interest. On the other hand, purposive sampling can also be conducted using a completely different strategy. In that case, a so-called heterogeneous sampling approach is applied (Qualtrics, n.d.). The goal here is to ascertain all the diversity that exists within the population. The VLingS project applied both strategies across various population levels. Groups of vulnerable language speakers were selected based on the principle of heterogeneous sampling, as the researchers aimed to include as many linguistic communities as possible whose languages had previously been assessed as vulnerable. This approach is logical, considering that the primary goal of the VLingS project was to create a questionnaire for assessing the vulnerability degree of certain languages and linguistic

varieties, an instrument that had not previously existed in our region. Conversely, heterogeneous sampling is often encountered precisely in the initial phases of research on various social phenomena (Fajgelj 2010), which is indeed the case here. In contrast to the selection of groups of vulnerable language speakers, the principle of selecting modal members was applied to select individual representatives from a given language group.

The application of both described strategies during the research was dictated by the researchers' goal to include as many speakers of a particular language or members of linguistic communities (if active speakers are unavailable) as possible once that language was characterised as vulnerable. This goal was adjusted considering the significant differences in the size of linguistic communities. In other words, in communities with just a few dozen respondents, such as Ladino speakers, the aim was to examine (almost) all speakers. In communities with a significantly larger number of members, such as speakers of Romani or Vlach, the objective was to examine segments of those subpopulations.

5. Quota sampling

The number of respondents from a specific language community was not determined arbitrarily but based on quotas. The quotas were defined based on official and unofficial estimates of the number of speakers of the target languages and their share in the total number of speakers of all vulnerable languages (see Table 1). The reason for relying on informal estimates of the number of members of certain language communities is that the census does not provide data on all groups of interest.

Quota sampling is defined by first selecting a characteristic of the population to be studied (e.g. gender), determining the proportion of all modalities of that characteristic (the share of men, women, non-binary individuals, etc.), and, based on this, forming quotas for the interviewers so that the proportion of modalities in the sample reflects the share of the same modalities in the entire population (Fajgelj 2010). In the example of gender, this would mean that if the ratio of men to women in a given country is 52% to 48%, a sample of 100 respondents would include 52 males and 48 females.

Within the VLingS project, the selection of respondents was carried out based on only one quota—the endangered language of the community to

which the individual belongs. This quota falls under the category of "hard" characteristics. More precisely, these are all characteristics that can be easily and objectively determined (Fajgelj 2010). The initial step in defining the quotas was to determine the number of speakers of a particular language. Given that official (census) and unofficial estimates varied significantly, it was impossible to precisely determine the share that the speakers of a specific language should have in the overall sample. Nevertheless, the quotas for the final sample were ultimately defined to approximately balance the information about the number of respondents from both sources.

When assessing language vulnerability, the age of respondents is an important factor. Documenting the age of speakers of a given language can facilitate tracking the extent of language transmission from older to younger generations. However, the age of respondents was not singled out as a specific quota. There were two reasons for this decision. First, to define a characteristic as a quota, it is necessary to know its distribution in the population (Fajgelj 2010). The 2011 census provides detailed data on the population's age structure (Statistical Office of the Republic of Serbia 2012). However, as previously mentioned, some speakers of the target languages were not registered in the census as a separate group, which prevents the assessment of age distribution. Additionally, estimates of the number of speakers of some languages vary significantly, contributing to the inability to specify the share of speakers of a certain age within a given language. It should also be noted that researchers estimated that there are fewer than 10 speakers of Ladino. This indicates that it is not even possible to speak of age group distribution in the true sense of the word, especially one that would be statistically useful. Furthermore, during the pilot study, no speakers of the Megleno-Romanian were found, so this language is now considered extinct in Serbia[2].

The second reason for excluding age as a quota was to avoid nested quotas. Nested quotas are characterised by the fact that one variable is "nested"

[2] This conclusion appears to contradict the information presented in Table 1, column VLingS pilot (realised). That column indicates that one speaker of Megleno-Romanian language was surveyed. The individual in question is someone whose parents spoke this language, a fact established approximately 15 years ago when Annemarie Sorescu-Marinković, one of the researchers on this project, interviewed them. For this reason, the individual was contacted during pilot research. However, he does not speak Megleno-Romanian, leading to the conclusion that this language no longer exists in Serbia. Hence the apparent inconsistency in the data.

within another (Fajgelj 2010). If nested quotas had been present in this research, the researchers would have been instructed to find a predetermined number of members of a language community and then, within each of them, find a specific number of respondents of different ages so that their share reflects the proportion of those age groups in the population. Such quotas significantly complicate the process of finding respondents (Fajgelj 2010). Initially, the interviewer is free to choose practically any person, typically those whose combination of characteristics is more common in a given population. Once they fill these quotas, they are left with those that have less common combinations of traits. This can sometimes be a difficult, if not impossible, task. In this research, such a task was unachievable in language communities that have a handful of members. The situation may not be any easier in communities with a higher number of respondents. For example, researchers' fieldwork experiences show that members of minority communities are not accustomed to participating in studies of this kind, and tend to avoid researchers when they appear. Specifically, there have been instances where potential participants openly refused the researcher or told them they would be home at a certain time but were not.

6. Snowball sampling

When selecting specific respondents, particularly in communities with a small number of speakers, the researchers also applied snowball sampling. This technique relies on an initial selection of individuals with certain characteristics, who then expand the sample by recommending other people with the same or similar traits (Milas 2009, Fajgelj 2010). Its application is particularly advisable when individuals with those characteristics are scarce, making it difficult to reach them. In such situations, it is generally expected that members of the population are in contact with one another (Hedrih & Hedrih 2022).

Field experience during the project's implementation shows that, in a number of cases, respondents recommended each other, family members or neighbours, for participation in the research. On the other hand, researchers were sometimes proactive and asked participants to recommend another person as a potential respondent. In these situations, they often had clearly defined requirements. For example, if the assessment indicated the need

to include a younger respondent, researchers would ask participants to recommend someone of a specific age. In some other communities, a slightly different tactic was applied. Specifically, the researchers who interviewed speakers of the Romani initially contacted "prominent" community members (e.g. presidents of local Romani associations or Romani language teachers), who then recommended others to be included in the sample. A similar tactic was applied by researchers studying speakers of the Rusyn language. They first contacted professors of that language from the Faculty of Philosophy at the University of Novi Sad, receiving recommendations on whom to interview next. According to the researchers' experience, this significantly facilitated establishing contacts with potential respondents. The fact that the researchers informed respondents that a specific person had already recommended them also positively affected their motivation to participate in the study, as did the information that a number of people from the community had already been interviewed.

The choice of snowball sampling, however, was also beneficial for the researchers working in communities that are not as small, such as the Vlach and Bayash Romanian communities. Since these rural communities are very traditional and patriarchal, a recommendation from a community member, often a member of the nuclear or extended family, carries more weight than recommendations from others. In some cases, such recommendations had an additional selection character. For example, before going to a specific place of residence, the researcher would have the names of a few people they had initially planned to contact. After the first conversation, the respondent would ask the researcher whom they planned to visit next. Upon hearing that the researcher intended to go to a specific person, the respondent would assess whether the next potential respondent had sufficient knowledge about the topic and would then offer their recommendation based on that—either the planned respondent or someone else about whom the researcher had not known beforehand.

7. Determining the number of respondents

As noted several times, one of the main goals of the VLingS project was to create a questionnaire to collect information on various aspects of endangered languages. Questionnaires, by nature, serve to gather data

on a large number of variables to assess and examine different aspects of the phenomenon of interest. Within this project, the plan was to subject aforementioned variables to statistical analysis. In order to draw valid statistical conclusions about the relationships between variables, it was necessary to estimate statistical power. The *statistical power* or *the power of a statistical test* refers to the probability that the null hypothesis (H0 – the hypothesis of no differences between populations) will be rejected when it is actually false (Faul et al. 2007, Milas 2009). In other words, it is the probability that the applied statistical procedure will show that there is a certain difference between groups or that two variables are correlated when such a difference or correlation truly exists in the entire population. This second hypothesis, which states that there is a difference between groups or a correlation between variables, is usually called the alternative hypothesis and is denoted as H1.

The chi-square test of independence can be used as a concrete example. This test is based on the null hypothesis that there is no association between two categorical variables. If this procedure yields statistically significant results, it means that we reject the null hypothesis and assert that there is sufficient evidence for an association between the two variables. It is also worth noting that each statistical test (chi-square, *t*-test, ANOVA, etc.) relies on its specific null hypothesis (Pallant 2016).

The power of a statistical test, i.e. its ability to show that the null hypothesis is incorrect when it is indeed incorrect, depends on several factors: the adopted significance level (the so-called alpha), the variance of the variables, the effect size (the size of the difference between groups or the degree of association between variables), the statistical test used, and the number of respondents to be included in the study (Milas 2009). The value of α, which relates to the significance level, also represents the probability that the researcher will make a Type I error. This error pertains to false positive results. If someone makes a Type I error, it means they have rejected the null hypothesis as false, even though it is actually true. In other words, it indicates that our statistical test shows that an effect exists, although in reality, it does not. If we want to increase the statistical power of the test, we will assume a greater risk of rejecting the correct null hypothesis, i.e. the value of α. However, in practice, the value of α is almost always set at .05, meaning that the researcher accepts a 5% risk of making a Type I error. Accordingly, a level of statistical significance of .01 indicates a 1% chance of making this error.

Moreover, the power of the design will be greater if the effect size is larger. In other words, the greater the real difference between groups of respondents or the real correlation between two variables in the population, the more easily it will be detected by a given statistical procedure. Finally, we can also influence the statistical power by increasing the number of respondents included in a particular study. This approach is considered the most common and straightforward way to increase statistical power (Milas 2009).

Tests that lack sufficient statistical power are unable to adequately differentiate between H0 and H1, which is the principal reason why it is necessary to conduct a statistical power analysis. Faul et al. (2007) state that statistical power depends on three classes of parameters: 1) significance level (α), 2) the number of respondents included in the study, and 3) effect size – a parameter based on which we define the alternative hypothesis (H1) as a certain degree of deviation from H0 in a given population.

There are various types of power analysis depending on the parameters that researchers have at their disposal and wish to assess. Since our goal was to determine the minimum number of respondents required for the application of appropriate statistical tests to yield usable results, an *A Priori Power Analysis* was conducted. As the name implies, a priori power analysis is carried out prior to the actual research, i.e. in the preparation phase. For this purpose, the specialised software GPower v3.1 was used.

A priori power analysis serves to determine the required number of respondents to be included in a given study based on the values of three pre-defined parameters: 1) the power of the design (which is determined as 1-β, where β refers to the Type II error, i.e. the decision to retain an incorrect H0); 2) the significance level (α); and 3) the population effect size that will be detected with the probability of 1-β. A priori analysis is an effective method for estimating statistical power before conducting the study, and its use is recommended in situations where time and financial resources for its implementation are not strictly limited (Faul et al. 2007).

We have already said that the value of α is usually taken to be .05, indicating a 5% chance of making a Type I error and a 95% chance that the obtained result is accurate. On the other hand, there is a consensus that the power (1-β) is set at a probability of at least .8. In other words, this means there is an equal or greater than 80% chance that the effect will be detected if it actually exists. Recently, this threshold was increased to .9 (Kiernan

& Baiocchi 2022), establishing a "stricter" criterion for detecting effects. This stricter criterion, i.e. greater statistical power, also means that a larger number of respondents will be needed to achieve statistically significant results at the same values of α and population effect size.

All the aforementioned facts were factored in when determining the necessary number of respondents for the VLingS project. In fact, the first step was to define the nature of the variables (nominal, ordinal, or interval) when forming the SPSS database. In the next step, the researchers defined potential research questions (problems) and, based on that, the statistical procedures to be used during data analysis were selected. Consequently, a series of procedures was generated within the GPower program to estimate the required number of respondents. The minimal number of respondents was estimated for the following procedures: ANOVA, chi-square, linear regression, Pearson's correlation coefficient, and point-biserial correlation coefficients. The input parameters were defined based on strict criteria. This means that effect sizes were in the range of values between .1 and .2, the α was set at .05, and the power (1-β) was raised to .95, which is above the "strict" convention that has recently been required (.9). In other words, we opted for an estimation of the number of respondents needed to register very low effects in the population, with a 5% risk of making a Type I error, but with a reduction of the risk of making a Type II error (to remain with the null hypothesis, although it is actually false) to 5%. This series of procedures produced an estimate of the number of respondents, ranging from 325 for the chi-square test to 1,293 for Pearson's correlation coefficient.

Based on the results of the described procedures, we decided that the final number of respondents to be included in the research would be 800. The number of respondents was adjusted considering the resources available to the researchers, the total estimated number of speakers of endangered languages, their alignment with the quotas, the geographical distribution of respondents, as well as the stringency of the criteria utilised in the a priori power analysis. This stringency allowed for a reduction in the number of respondents without affecting the analyses.

Let us take Pearson's correlation coefficient as an example and assess the number of respondents based on the "strict" values of input parameters (Figure 1). Here, we see that to detect a small effect size (p H1 = .1), at a significance level (α) of .05 and a power of .95, 1,293 respondents are needed. However, when we decrease the power to .8, which aligns with the generally

accepted convention, the required number of respondents drops to 782 (Figure 2). On the other hand, if we define the power according to the new convention of .9 and increase the effect size by only .05 (so it now equals .15, which is still a small effect), the required number of respondents falls to 462.

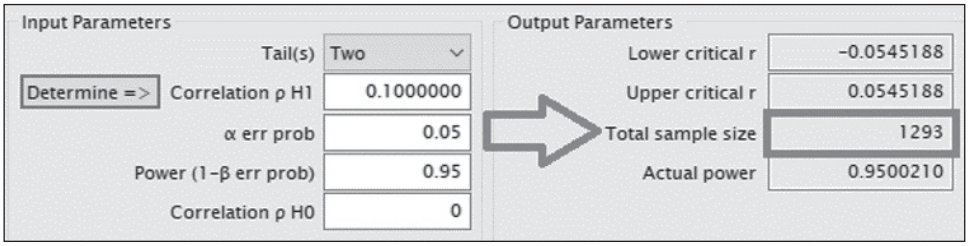

Figure 1: Estimation of the number of respondents for Pearson's correlation coefficient – stricter criteria.

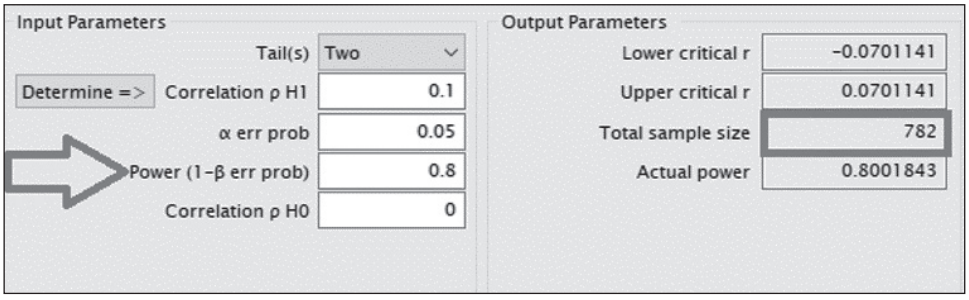

Figure 2: Estimation of the number of respondents for Pearson's correlation coefficient – less strict criteria.

Similar results are obtained when the parameters in the a priori power analysis for other mentioned statistical tests are adjusted in the same way. For this reason, the number of respondents could be adjusted to meet the practical demands of the research without compromising its ability to identify associations between variables if they truly exist at the population level. Likewise, this allowed for potential deviations from the planned 800 respondents if the field researchers were unable to interview so many individuals.

The number of 800 respondents refers to the total number of individuals in the sample. The number of speakers from a specific language community was determined based on quotas. As previously mentioned, these numbers

were set so that their ratio in the sample reflects the ratio that exists in the entire population of speakers of endangered languages. The total number of speakers was defined based on official and unofficial estimates of their numbers in the population of the Republic of Serbia. The values of quotas are presented in Table 1.

When all the aforementioned considerations are taken into account, particularly the strict criteria for defining the input parameters in the a priori power analysis, it can be concluded that the final number of speakers included in the study (n=686) did not compromise the power of the applied statistical tests. In other words, the final number of respondents is sufficient for various tests to detect very small effects if they truly exist.

8. Conclusion

The selection of the type and size of the sample(s) was just one aspect of the methodological preparation done for the study of endangered and vulnerable languages in Serbia. However, this presented portion of the preparation sufficiently illustrates of the complexity involved in researching certain social phenomena. The purpose of explaining the reasons for choosing non-probability samples, as well as the need to combine several different types of samples, is to highlight the specificity and multidimensionality of these studies. The essence of the problems we encountered at the beginning of our research is presented in Table 1, which serves as another representation of the complexities of social reality—this time, the reality and necessity of investigating linguistic communities. To put it succinctly, this complexity stemmed from multiple factors, starting from the fact that researchers failed to find a single speaker of Megleno-Romanian, to estimates regarding the number of speakers based on prior research experience, and finally to census data. By combining purposive sampling (along with the strategy of selecting modal respondents and heterogeneous sampling), quota sampling, and snowball sampling, we addressed all specific methodological challenges of the sample. Recognizing that our study is not the only one where researchers have applied a combination of different sampling methods, our intention is to demonstrate to the academic community how we approached our research and the reasons that guided our choice of sampling methods.

On the other hand, we have explained to the readers the process of determining the total number of respondents, as well as all the factors that contributed to this determination. We believe this information is essential and serves to substantiate and enhance the reliability of the results obtained during data collection and subsequent statistical analysis. Thorough reflection and analysis of the parameters and criteria based on which the number of respondents in specific groups was defined contribute to the objectivity of the knowledge we have gained by studying endangered and vulnerable languages in Serbia.

References

Anderson, J. A., L. Mak, A. Keyvani Chahi and E. Bialystok. 2018. "The language and social background questionnaire: Assessing degree of bilingualism in a diverse population". *Behavior Research Methods* 50, 250–263.

Brenzinger, M., A. M. Dwyer, T. de Graaf, C. Grinevald, M. Krauss, O. Miyaoka, N. Ostler, O. Sakiyama, M. E. Villalón, A. Y. Yamamoto and O. Zapeda. 2003. "Language vitality and endangerment". Document submitted to *the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages*, Paris, 10–12 March 2003. Available at: https://ich.unesco.org/doc/src/00120-EN.pdf. Accessed August 29, 2024.

Đurić, V., D. Tanasković, D. Vukmirović and P. Lađević. 2014. *Etnokonfesionalni i jezički mozaik Srbije* [The Ethno-Confessional and Linguistic Mosaic of Serbia]. Beograd: Republički zavod za statistiku.

Fajgelj, S. 2010. *Metode istraživanja ponašanja* [Behavioral Research Methods]. Beograd: Centar za primenjenu psihologiju.

Faul, F., E. Erdfelder, A. G. Lang and A. Buchner. 2007. "G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences". *Behavior Research Methods* 39 (2), 175–191.

Fishman, J. A. 1991. *Reversing Language Shift. Theoretical and Empirical Foundations of Assistance to Threatened Languages.* Clevedon: Multilingual Matters Ltd.

Hedrih, V. and A. Hedrih. 2022. *Osnove psihološke statistike* [Fundamentals of Psychological Statistics]. Niš: Filozofski fakultet.

Kiernan, M. and M. T. Baiocchi. 2022. "Casting new light on statistical power: an illuminating analogy and strategies to avoid underpowered trials". *American Journal of Epidemiology* 191 (8), 1500–1507.

Krauss, M. 2008. "Chapter 1. Classification and Terminology for Degrees of Language Endangerment". In *Language Diversity Endangered*, ed. M. Brenzinger, 1–8. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110905694-003

Lee, N. H. and J. van Way. 2016. "Assessing levels of endangerment in the Catalogue of Endangered Languages (ELCat) using the Language Endangerment Index (LEI)". *Language in Society* 45, 271–292.

Li, P., S. Sepanski and X. Zhao. 2006. "Language history questionnaire: A web-based interface for bilingual research". *Behavior Research Methods* 38 (2), 202–210. https://doi.org/10.3758/BF03192770

Lewis, M. P. and G. F. Simons 2010. "Assessing endangerment: Expanding Fishman's GIDS". *Revue Roumaine de Linguistique* 55 (2), 103–120.

Marian, V., H. K. Blumenfeld and M. Kaushanskaya. 2007. "The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals". *Journal of Speech, Language, and Hearing Research* 50 (4), 940–967.

Milas, G. 2009. Istraživačke metode u psihologiji i drugim društvenim znanostima [Research Methods in Psychology and Other Social Sciences]. Zagreb: Naklada Slap.

Mirić, M., V. Sokolovska and A. Sorescu-Marinković. 2024. "*Mother tongue* in Serbia: A speakers' perspective on the meaning of the concept". *Open Linguistics* 10 (1). https://doi.org/10.1515/opli-2024-0016

Pallant, J. 2016. *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. London: McGraw-Hill Education.

Qualtrics. n.d. *What is non-probability sampling? Everything you need to know*. Available at: https://www.qualtrics.com/experience-management/research/non-probability-sampling/. Accessed October 22, 2024

Radić-Bojanić, B. and V. Sokolovska. 2024. "Challenges in Researching Endangered and Vulnerable Languages in Serbia within the Project "Vulnerable Languages and Linguistic Varieties in Serbia": Methodological and Ethical Issues". *Etnoantropološki problemi/Issues in Ethnology and Anthropology* 19 (1), 279–308.

Sorescu-Marinković, A., M. Mirić and S. Ćirković. 2020. "Assessing linguistic vulnerability and endangerment in Serbia a critical survey of methodologies and outcomes". *Balcanica* LI, 65–104.

Statistical Office of the Republic of Serbia. 2012. *2011 Census of Population, Households and Dwellings in the Republic of Serbia: Age and Sex Data by Settlements*. Available at: https://publikacije.stat.gov.rs/G2012/Pdf/G20124002.pdf. Accessed December 15, 2024.

Acknowledgement

The manuscript is the result of the project "Vulnerable Languages and Linguistic Varieties in Serbia" (VLingS), funded by the Science Fund of the Republic of Serbia (grant number: 7736100) within the program IDEAS (2022–2024). The funding was provided through the Budget of the Republic of Serbia, and the World Bank project – the Serbia Accelerating Innovation and Entrepreneurship Project (SAIGE).