MA ADELAJDA BAFTIU

**REZYME**

Punimi i doktoratës “Korpusi i folur i shqipes në Kosovë” përbëhet nga gjashtë kapituj. Në Hyrje trajtohen objekti i studimit, metoda e studimit, hipoteza, rëndësia dhe detyrat e studimit. Objekt i këtij studimi është krijimi i një aparature shkencore për përpilimin e korpusit të folur në Kosovë. Ky studim i përket fushës së gjuhësisë së korpusit dhe është nisur duke u mbështetur në hipotezën se korpusi i mbledhur në këtë punim do të jetë përfaqësues i shqipes së folur në Kosovë dhe do të përfaqësojë varietetet rajonale dhe shoqërore të shqipes së folur në Kosovë. Këtë hipotezë jemi përpjekur ta vërtetojmë përmes metodës statistikore dhe përmes qasjeve kryesore që përdoren për krijimin e korpusit. Dy qasjet kryesore që përdoren në gjuhësinë e korpusit janë qasja e bazuar në korpus (*The corpus based approach*) dhe qasja e drejtuar nga korpusi (*The corpus driven approach*). Në këtë studim do të përdoret qasja e bazuar në korpus, sepse shembujt që dalin nga korpusi synohen të jenë p

ërgjithësues për shqipen e folur në Kosovë. Qëllimi i këtij studimi është mbledhja dhe përpilimi i korpusit të folur. Në hulumtim përfshihen të gjitha regjionet e Kosovës. Numri i intervistave në secilin qytet është vendosur duke u bazuar në numrin e banorëve të asaj komune dhe në numrin e fshatrave, p.sh. nëse Prishtina i ka 208.230 banorë dhe 42 fshatra, në këtë komunë janë mbledhur afërsisht 200 intervista, 1 intervistë në 1000 banorë, pra në secilin fshat së paku nga 2-3 intervista. Ky korpus synon t’i përmbajë 500.000 fjalë, pra 2.000 incizime/mostra, secila mostër përafërsisht të ketë 250 fjalë. Kohëzgjatja mesatare e incizimeve është 6 minuta.

Mbledhja e intervistave ka filluar në vitin 2019 dhe ka përfunduar në vitin 2021. Shumica e incizimeve janë kryer përmes intervistave joformale, me pyetje të përgatitura në varietetin e shqipes së folur në Kosovë, me qëllim që edhe folësit të përgjigjen në të njëjtin varietet. Mjedisi apo vendi ku janë kryer pjesa më e madhe e intervistave ka qenë nëpër shtëpi, shkolla, universitete, dyqane, sheshe, kafene, zyra etj. Pas intervistës, është plotësuar anketa nga subjektet me këto të dhëna: vendlindja, vendbanimi, mosha, gjinia, profesioni, shkollimi, gjuha amtare dhe gjuhët e huaja. Subjektet u përkasin profesioneve të ndryshme: mësues, tregtarë, ekonomistë, profesorë, kuzhinierë, inxhinierë, artistë, infermierë, sociologë, përkthyes, policë, etj. Te grupmosha e vjetër ka raste edhe pa shkollim fare, ndërsa te grupmosha e mesme dhe e re dalin tri nivelet e shkollimit: i ulët, i mesëm dhe i lartë. Pëveç këtyre që u përmendën, subjektet janë marrë në mënyrë të barabartë nga gjinia femërore e mashkullore dhe janë afërsisht të moshës 18-90 vjeç.

Shumica e subjekteve si gjuhë amtare e kanë gjuhën shqipe, por flasin edhe gjuhë të huaja, si: anglishten, gjermanishten, frëngjishten, turqishten, spanjishten dhe serbishten. Secila anketë pastaj është ruajtur edhe në format elektronik me të njëjtin emërtim si incizimi, p.sh. *Prishtinë, 1, Prishtinë, 2* etj.

Në kapitullin e parë flitet për gjuhësinë e korpusit. Analiza e të dhënave gjuhësore mbështetur në koleksione tekstesh të folura ose të shkruara që ruhen në format elektronik njihet si gjuhësi korpusi. Disa autorë gjuhësinë e korpusit e konsiderojnë si disiplinë më vete, ndërse disa të tjerë thjesht si metodologji. Sipas McEnery dhe Wilson (1996), gjuhësia e korpusit nuk është fushë e gjuhësisë e krahasueshme me sintaksën, semantikën, sociolinguistikën, sepse këto disiplina koncentrohen në përshkrimin e disa aspekteve të përdorimit të gjuhës, ndërsa gjuhësia e korpusit është më shumë metodologji. Të njëjtin mendim ndan edhe Kennedy (1998), megjithëse thekson se gjuhësia e korpusit shkon përtej përdorimit të saj si burim provash në përshkrimin gjuhësor. Në anën tjetër, Leech (1992) e definon si një qasje të re kërkimore e filozofike ndaj temës. Gjuhësinë e korpusit si disiplinë e shohin autorë si Tognini-Bonelli (2001), ajo thekson se gjuhësia e korpusit shkon përtej rolit të pastër metodologjik. Edhe McEnery dhe Hardie (2012) në studime të mëvonshme thonë se gjuhësia e korpusit e ka potencialin për ta riorientuar tërë qasjen tonë për studimin e gjuhës dhe mund të rafinojë e të ripërcaktojë një varg teorish të gjuhës, pra është një fushë heterogjene. Baker (2010: 1) po ashtu thekson rëndësinë që ka pasur gjuhësia e korpusit në njëzet vitet e fundit si një qasje për studimin e gjuhës dhe pranimin si një mënyrë e rëndësishme dhe e dobishme e kërkimit gjuhësor. Pra, mund të pranohet se gjuhësia e korpusit shkon përtej rolit të saj si metodologji.

Kapitulli i parë ndahet në pesë nënkapituj: korpusi, historiku i gjuhësisë së korpusit, historiku i gjuhësisë së korpusit në gjuhën shqipe, llojet e korpusit dhe gjuhësia e korpusit dhe sociolinguistika. Në nënkapitullin e parë është trajtuar çfarë tipa tekstesh përbëjnë një korpus dhe cilat janë kriteret që një tekst të quhet korpus. Në nënkapitullin e dytë është trajtuar historiku i gjuhësisë së korpusit në përgjithësi. Korpusi i Brownit është korpusi i parë kompjuterik i përpiluar për qëllime gjuhësore nga Nelson Francis dhe Henry Kučera (1964). Në nënkapitullin e tretë është trajtuar historiku i gjuhësisë së korpusit në gjuhën shqipe. Në Kosovë, pas vitit 2000, jepen vërejtjet e para për nevojën e krijimit të një korpusi në gjuhën shqipe. Nënkapitulli i katërt përmbledh llojet e korpusit: korpusi gjeneral, korpusi krahasues, korpusi i specializuar, korpusi historik, korpusi për qëllime të veçanta, korpusi shumëgjuhësor, korpusi i folur, korpusi i analizuar gramatikisht,

korpusi multimedial, korpusi ueb dhe korpusi *lingua franca.* Në nënkapitullin e pestë trajtohet lidhja në mes të gjuhësisë së korpusit dhe sociolinguistikës.

Në kapitullin e dytë është folur për krijimin e korpusit. Për krijimin e korpusit është shumë e rëndësishme që pyetja kërkimore të jetë e qartë. Varësisht prej pyetjes kërkimore, përcaktohet edhe madhësia e korpusit, dizajni i korpusit dhe mbledhja e teksteve. Në këtë kapitull është folur edhe për mjetet për shënimin dhe transkriptimin e korpusit.

Në kapitullin e tretë është folur për varietetet e gjuhës shqipe. Ismajli (2019) thekson se historikisht brenda shqipes janë zhvilluar dy ndarje të mëdha dialektore: gegnishtja në veri dhe toskërishtja në jug të lumit Shkumbin. Toskërishtja flitet në jug nga Struga deri tek Adriatiku në nivel të Shkumbinit, duke e përfshirë edhe shqipen e folur në Greqi dhe në Itali, ndërsa gegnishtja flitet në veri, duke përfshirë edhe arbneshët e Zarës. Kjo ndarje vazhdon tutje në: gegnishten jugore, gegnishten qendrore, gegnishten veriperëndimore/verilindore, ndërsa në jug toskërishten veriore, labërishten dhe çamërishten dhe arvanitishten e arbërishten e Italisë. Ismajli (2019) më tutje vëren se izoglosat që shërbejnë për ndarje të këtilla nuk i përkasin të njëjtës periudhë historike. Ndarja e dialekteve dytësore vijon në: arvanitika, arbërishte dhe në arbnishte (Ismajli 2019: 347-348). Në këtë kapitull janë përmbledhur edhe veçori gjuhësore të të folmeve të ndryshme të Kosovës, si nga: e folmja e Deçanit, e folmja e Rugovës, e folmja e Hasit, e folmja e Karadakut, e folmja e Shalës së Bajgorës, e folmja e Gjakovës, e folmja e Kaçanikut, e folmja e Istogut, e folmja e Opojës, e folmja e Rahovecit etj. Në pjesën e fundit të këtij kapitulli është diskutuar edhe për pozitën dhe statusin e shqipes në vendet shqipfolëse dhe për kontaktet e shqipes.

Në kapitullin e katërt është përshkruar dizajni i korpusit të folur në Kosovë. Ky kapitull ndahet në gjashtë nënkapituj: dizajni i korpusit të folur nëpër komunat e Kosovës (vendlindja dhe vendbanimi), mosha e të anketuarve, gjinia e të anketuarve, shkollimi i të anketuarve, profesionet e të anketuarve dhe gjuhët e dyta që fliten në Kosovë. Në gjithë Kosovën janë realizuar 1800 intervista. Numri i ndryshëm i intervistave nëpër komuna del për shkak të madhësisë së ndryshme demografike, por edhe numrit të ndryshëm të vendbanimeve në kuadër të komunave. Në Komunën e Deçanit janë realizuar 114 intervista, në Komunën e Gjakovës 113, në Komunën e Gllogocit 18, në Komunën e Gjilanit 121, në Komunën e Dragashit 17, në Komunën e Istogut 25, në Komunën e Kaçanikut 47, në Komunën e Klinës 10, në Komunën e Fushë Kosovës 20, në Komunën e Kamenicës 2, në Komunën e Mitrovicës 163, në Komunën e Lipjanit 114, në Komunën e

Novobërdës 2, në Komunën e Obiliqit 20, në Komunën e Rahovecit 125, në Komunën e Pejës 131, në Komunën e Podujevës 133, në Komunën e Prishtinës 117, në Komunën e Prizrenit 35, në Komunën e Skënderajt 22, në Komunën e Shtimes 93, në Komunën e Shtërpcës 21, në Komunën e Suharekës 29, në Komunën e Ferizajt 117, në Komunën e Vitisë 8, në Komunën e Vushtrrisë 28, në Komunën e Malishevës 113, në Komunën e Junikut 7, në Komunën e Hanit të Elezit 24 dhe në Komunën e Graçanicës 11 intervista. Përveç vendlindjes, vendbanimit, shkollimit, profesionit, gjuhës amtare dhe gjuhës së dytë, të anketuarit te të dhënat demografike kanë plotësuar edhe gjininë dhe moshën e tyre. Nga 1800 të anketuar, 787 subjekte i përkasin grupmoshës së re (18-39 vjeç), 442 grupmoshës së mesme (40-59 vjeç), 434 grupmoshës së vjetër (60-79 vjeç) dhe 137 subjekte nuk e kanë deklaruar moshën e tyre. Nga 1800 të anketuar, 844 subjekte janë të gjinisë femërore dhe 956 subjekte të gjinisë mashkullore. Shkollimi është ndarë në shkollimin e ulët, të mesëm dhe të lartë. Nga 1800 të anketuar, të pashkolluar janë 19 subjekte, 265 të anketuar e kanë shkollimin e ulët, 505 të anketuar e kanë shkollimin e mesëm, 626 subjekte e kanë shkollimin e lartë dhe 132 nuk e kanë deklaruar shkollimin e tyre. Nga 1800 të anketuar, 213 të anketuar janë të papunë, 232 janë studentë, 166 janë nxënës, 113 janë të pensionuar, 135 nuk e kanë deklaruar profesionin e tyre dhe 941 të anketuar kanë profesione, si: mësimdhënës, ekonomistë, juristë, inxhinierë, tregtarë, bujq, psikologë, punëtorë krahu etj. Gjuhët e dyta që fliten nëpër komunat e Kosovës janë: anglishtja, gjermanishtja, serbishtja, kroatishtja, sllovenishtja, rusishtja, boshnjakishtja, maqedonishtja, polonishtja, bullgarishtja, turqishtja, frëngjishtja, arabishtja, spanjishtja, italishtja, suedishtja, latinishtja, holandishtja, norvegjishtja, rumunishtja, hungarishtja, gjuha rome dhe greqishtja.

Në kapitullin e pestë jepen përfundimet nga rezultatet e hulumtimit.

Fjalët kyçe: Gjuhësi korpusi, korpus i folur, varietete të shqipes, komuna të Kosovës, sociolinguistikë

MA ADELAJDA BAFTIU

**SUMMARY**

The doctoral thesis "Albanian spoken corpus in Kosovo" consists of six chapters. The Introduction deals with the object of the study, the method of the study, the hypothesis, the importance and the tasks of the study. The object of this study is the compilation of the spoken corpus in Kosovo. This study belongs to the field of corpus linguistics and it started based on the hypothesis that the corpus collected in this paper will be representative of the spoken Albanian in Kosovo and will represent the regional varieties of spoken Albanian in Kosovo. We have tried to prove this hypothesis through the statistical method and through the main approaches used to create the corpus. The two main approaches used in corpus linguistics are *The corpus-based approach* and *The corpus-driven approach*. In this study, the corpus-based approach will be used, because the examples that emerge are intended to be general for spoken Albanian in Kosovo. The purpose of this study is to collect and compile the spoken corpus. All regions of Kosovo are included in the research. The number of interviews in each city is decided based on the number of inhabitants of that municipality and the number of villages, e.g. Prishtina has 208,230 inhabitants and 42 villages, in this municipality it was necessary to collect aproximately 200 interviews, 1 interview per 1000 inhabitants, so in each village at least 2-3 interviews. This corpus aims to contain 500,000 words, ie 2,000 recordings / samples, each sample approximately 250 words, with a duration of 6 minutes.

The collection of intervieës started in 2019 and ended in 2021. Most of the recordings were made through informal interviews, with questions prepared in the variety of spoken Albanian in Kosovo, in order for the speakers to answer in the same way. The environment or place where most of the interviews were conducted was in homes, schools, universities, shops, squares, cafes, offices, cars, etc. After the interview, the survey was completed by the subjects with the following data: place of birth, place of residence, age, gender, profession, education, mother tongue and foreign languages. Subjects belong to different professions: teachers, students, pupils, businessmen, economists, professors, cooks, engineers, artists, nurses, sociologists, translators, police officers, etc. In the old age group there are cases without education at all, while in the middle and young age group there are three levels of education: low, middle and high. In addition to those mentioned, the subjects were taken equally by females and males and are approximately aged 18-90 years. Most of the subjects have Albanian as their mother tongue, but they also speak English,

German, French, Turkish, Spanish and Serbian. Each survey was saved in electronic format with the same name as the recording, e.g. *Prishtina, 1*, *Prishtina, 2* etc.

The first chapter talks about corpus linguistics. The analysis of linguistic data based on collections of spoken or written texts that are stored in electronic format is known as corpus linguistics. Some authors consider corpus linguistics as a separate discipline, while others simply as a methodology. According to McEnery and Wilson (1996), corpus linguistics is not a field of linguistics comparable to syntax, semantics, sociolinguistics, because these disciplines concentrate on describing some aspects of language use, while corpus linguistics is more of a methodology. Kennedy (1998) shares the same opinion, although he emphasizes that corpus linguistics goes beyond its use as a source of evidence in linguistic description. While Leech (1992) defines it as a new research and philosophical approach to the subject. Authors such as Tognini-Bonelli (2001) see corpus linguistics as a discipline, she emphasizes that corpus linguistics goes beyond a pure methodological role. Even McEnery and Hardie (2012) in later studies say that corpus linguistics has the potential to reorient our entire approach to the study of language and can refine and redefine a range of theories of language, so it is a heterogeneous field. Baker (2010: 1) also emphasizes the importance that corpus linguistics has had in the last twenty years as an approach to the study of language and its acceptance as an important and useful way of linguistic research. So it can be acknowledged that corpus linguistics goes beyond its role as a methodology.

The first chapter is divided into five subchapters: the corpus, the history of corpus linguistics, the history of corpus linguistics in the Albanian language, types of corpus and corpus linguistics and sociolinguistics. In the first sub-chapter, what types of texts make up a corpus and what are the criteria for a text to be called a corpus are discussed. In the second sub-chapter, the history of corpus linguistics in general is discussed. Brown's corpus is the first computerized corpus compiled for linguistic purposes by Nelson Francis and Henry Kučera (1964). In the third sub-chapter, the history of corpus linguistics in the Albanian language is discussed. In Kosovo, after the year 2000, the first remarks were made about the need to create a corpus in the Albanian language. Subchapter fourth summarizes the types of corpus: general corpus, comparative corpus, specialized corpus, historical corpus, special purpose corpus, multilingual corpus, spoken corpus, grammatically analyzed corpus, multimedia corpus, web corpus and lingua franca corpus. The fifth subchapter deals with the connection between corpus linguistics and sociolinguistics.

The second chapter talks about the creation of the corpus. For the creation of the corpus it is very important that the research question is clear. Depending on the research question, the size of the corpus, the design of the corpus and the collection of texts are also determined. This chapter also talks about the tools for annotating and transcribing the corpus.

The third chapter talks about the varieties of the Albanian language. Ismajli (2019) points out that historically two major dialectal divisions have developed within Albanian:Gheg in the north and Tosk in the south of the Shkumbin River. Tosk is spoken in the south from Struga to the Adriatic at the level of Shkumbin, including the Albanian spoken in Greece and Italy, while Gheg is spoken in the north, including the Albanians of Zara. This division continues further into: southern Gheg, central Gheg, northwest/northeast Gheg, while in the south the northern Tosk, the lab variety and the *çamë* variety and the Arvanitika and Arberishte of Italy. Ismajli (2019) further notes that the isoglosses that serve for such divisions do not belong to the same historical period. The secondary dialects are divided into: Arvanitika, Arberishte and Arbënishte (Ismajli 2019: 347-348). In this chapter, the linguistic features of the various dialects of Kosovo are summarized, such as from: the variety of Deçani, the variety of Rugova, the variety of Has, the variety of Karadak, the variety of Shala e Bajgora, the variety of Gjakova, the variety of Kaçanik, the variety of Istog, the variety of Opoja, the variety of Rahovec, etc. The last part of this chapter also discusses the position and status of Albanian in Albanian-speaking countries and the contacts of Albanian.

In the fourth chapter, the design of the corpus spoken in Kosovo is described. 1800 interviews were conducted throughout Kosovo. In the fourth chapter is the design of the corpus spoken in Kosovo. This chapter is divided into six sub-chapters: the design of the corpus spoken in the municipalities of Kosovo, the age of the respondents, the gender of the respondents, the education of the respondents, the profession of the respondents and the second languages spoken in Kosovo. The different number of interviews in the municipalities is due to the different demographic size, but also the different number of settlements within the municipalities. In the municipality of Deçan, 114 intervieës were conducted, in the municipality of Gjakova 113, in the municipality of Gllogoci 18, in the municipality of Gjilan 121, in the municipality of Gragashi 17, in the municipality of Istog 25, in the municipality of Kaçanik 47, in the municipality of Klina 10 , in the municipality of Fushë Kosova 20, in the municipality of Kamenica 2, in the municipality of Mitrovica 163, in the municipality of Lipjan 114, in the municipality of Novobërda 2, in the municipality of Obiliq 20, in the municipality of Rahovec 125, in the municipality of Peja 131, in

the municipality of Podujeva 133, in the municipality of Pristina 117, in the municipality of Prizren 35, in the municipality of Skenderaj 22, in the municipality of Shtime 93, in the municipality of Shtërpce 21, in the municipality of Suhareka 29, in the municipality of Ferizaj 117, in the municipality of Vitia 8, in the municipality of Vushtrri 28, in the municipality of Malisheva 113, in the municipality of Junik 7, in the municipality of Han i Elez 24 and in the municipality of Graçanica 11 interviews. In addition to the place of birth, place of residence, education, profession, mother tongue and second language, the respondents in the demographic data have also filled in their gender and age. Out of 1800 respondents, 787 subjects belong to the young age group (18-39 years), 442 to the middle age group (40-59 years), 434 to the old age group (60-79 years) and 137 subjects did not declare their age. Out of 1800 respondents, 844 subjects are female and 956 subjects are male. Education is divided into primary, secondary and tertiary education. Of the 1800 respondents, 19 subjects are uneducated, 265 respondents have low education, 505 respondents have secondary education, 626 subjects have high education and 132 have not declared their education. Out of 1800 respondents, 213 respondents are unemployed, 232 are students, 166 are students, 113 are retired, 135 have not declared their profession and 941 respondents have professions such as: teachers, economists, lawyers, engineers, merchants, farmers, psychologists, manual ëorkers, etc. The second languages spoken in the municipalities of Kosovo are: English, German, Serbian, Croatian, Slovenian, Russian, Bosnian, Macedonian, Polish, Bulgarian, Turkish, French, Arabic, Spanish, Italian, Swedish, Latin, Dutch, Norëegian, Romanian , Hungarian, Romani and Greek.

In the fifth chapter, the conclusions from the research results are given.

Key words: Corpus languistics, spoken corpus, varieties of Albanian, municipalities of Kosovo, sociolinguistics