Acta Linguistica Petropolitana. 2020. Vol. 16.2. P. 275–305
DOI 10.30842/alp2306573716210

Albanian dialects in the light of language contact: A quantitative study of loanwords*

M. S. Morozova
Institute for Linguistic Studies, Russian Academy of Sciences, St. Petersburg;
morozovamaria86@gmail.com

M. A. Ovsjannikova
Institute for Linguistic Studies, Russian Academy of Sciences, St. Petersburg;
masha.ovsjannikova@gmail.com

A. Y. Rusakov
Institute for Linguistic Studies, Russian Academy of Sciences, St. Petersburg;
ayurusakov@gmail.com

Abstract. The article presents the results of a quantitative study of lexical borrowings in the Albanian dialects based on the material from the Dialectological Atlas of the Albanian language (DAAL). This study is a stage in the investigation of Albanian dialects by the methods of areal typology and dialectometry aiming to clarify the existing ideas about the development of the Albanian dialectal landscape and to reconstruct the contact history of the Albanian language area.

The article analyzes the overall sample of lexical borrowings and the borrowings of different origins (Balkan Slavic, Medieval and Modern Greek, Ottoman Turkish, Eastern and Western Romance) found in Albanian dialectal varieties. Within the Albanian language area, we identify zones more or less affected by borrowing, and microareas characterized by various degrees of contact, isolation, exposure to the general Albanian linguistic development tendencies, etc. Using distance calculation and multidimensional scaling, we measure and map the closeness of Albanian varieties based on certain groups of borrowings and verify the existing views on the Albanian dialect classification and on the areal distribution of loanwords across the traditionally defined zones of Albanian-Slavic, Albanian-Romance, and Albanian-Greek contact.

Our quantitative analysis of the closeness of Albanian varieties shows that the results based on the overall sample of borrowings better correspond to the traditional dialectal classification than those based on any of the specific subgroups of borrowings. Some long-established subdialects of Gheg and Tosk demonstrate lack of internal homogeneity.

* The research was supported by the Russian Science Foundation (Grant No. 19-18-00244 “Balkan bilingualism in dominant and equilibrium contact situation in diatopy, diachrony and diastraty”).

Central Gheg varieties of North Macedonia, for example, clearly stand out in the Central Gheg subdialect, and tentative subdivisions are evident within Northern and Southern Tosk in the Tosk dialect area.

The quantitative distribution of loanwords shows a clear areal pattern, with the intensity of borrowing (and language contact) increasing from the center to the periphery of the Albanian-speaking area. While certain micro-areas differ in the number of borrowings from Slavic, Greek, or Arumanian lexis, no clear areal patterns are observable for the distribution of Turkish loanwords. As to Western Romance loanwords, relatively high numbers of these are not only, quite expectedly, found in the coastal Northwestern and Central Gheg varieties, but also in the most isolated Central Gheg varieties — a fact that may throw light on the early history of these dialects.

Keywords: Albanian, dialect, Gheg, Tosk, loanwords, Balkan Slavic, Ottoman Turkish, Greek, Eastern (Balkan) Romance, Western Romance, quantitative analysis, closeness, language contact.

Албанские диалекты в свете языкового контакта: количественное исследование заимствований

М. С. Морозова

Институт лингвистических исследований РАН, Санкт-Петербург;
morozovamaria86@gmail.com

М. А. Овсянникова

Институт лингвистических исследований РАН, Санкт-Петербург;
masha.ovsjannikova@gmail.com

А. Ю. Русаков

Институт лингвистических исследований РАН, Санкт-Петербург
ayurusakov@gmail.com

Аннотация. В статье представлены результаты количественного исследования лексических заимствований в албанских диалектах, основанного на материале Диалектологического атласа албанского языка (ДААЯ). Это исследование является одним из этапов изучения албанских диалектов методами, используемыми в ареальной типологии и диалектометрии, с тем чтобы верифицировать существующие представления о формировании албанского диалектного ландшафта и реконструировать контактную историю албаноязычного ареала.

В статье анализируются лексические заимствования различного происхождения (из балканославянских языков, средневекового греческого и новогреческого, османского турецкого, западнороманских и восточнороманских языков), которые

обнаруживаются в албанских говорах. Мы выявляем внутри албаноязычного ареала зоны с большей или меньшей интенсивностью заимствования лексики, и микрорегионы, которые характеризуются разной степенью интенсивности контакта, уровнем изолированности, проявлением общеалбанских тенденций языкового развития и др. Путем подсчета расстояний и последующего анализа и визуализации данных методом многомерного шкалирования мы измеряем и отражаем на карте степень близости албанских говоров, основанной на присутствии в них тех или иных групп заимствований, и предпринимаем попытку верифицировать существующие представления об албанской диалектной классификации и об ареальной дистрибуции заимствований в традиционно выделяемых зонах албанско-славянского, албанско-романского и албанско-греческого контакта.

Количественный анализ близости албанских говоров показал, что результаты, полученные на основе выборки из всех заимствований, более точно соответствуют общепринятой диалектной классификации, чем данные о близости говоров, основанные на анализе отдельных групп заимствований. Некоторые группы говоров отличаются внутренней неоднородностью. Например, в среднегегской диалектной зоне четко выделяются среднегегские говоры Северной Македонии. Различия обнаруживаются и между отдельными районами внутри севернотоскского и южнотоскского ареалов.

В распределении заимствований наблюдается хорошо различимая тенденция к повышению интенсивности заимствования (и языкового контакта) в говорах, расположенных на периферии албаноязычного ареала. Выделяются отдельные микроареалы с большим или меньшим числом заимствований разного происхождения — славянских, греческих, арумынских. В дистрибуции турцизмов, напротив, не удалось выявить четкого ареального распределения. Заимствования из западнороманских языков сравнительно многочисленны в прибрежных среднегегских и северо-западных гегских говоров, но также и в наиболее изолированных среднегегских говорах материковой части Албании, что позволяет пролить свет на раннюю историю этих диалектных групп.

Ключевые слова: албанский язык, диалект, гегский, тоскский, балканославянский, османский турецкий, греческий, восточнороманский (балканороманский), западнороманский, количественный анализ, близость, языковой контакт.

1. Introduction

Lexical borrowings, or loanwords, are recognized as the most commonly attested language contact phenomena. One of the questions arising in a study of loanwords is what their spatial and quantitative distribution in a given language tells us about the history, intensity, and setting of language contact in the area where it is spoken.

This article examines loanwords in Albanian dialects against a background of the "contact history" of Albanian. This study is a part of a larger research project investigating the Albanian dialect continuum by quantitative methods (see [Rusakov, Morozova 2017, 2018] on the linguistic complexity and closeness of Albanian dialectal varieties based on the grammatical features; [Rusakov et al. 2018] on the closeness based on the lexicon). Our goal here is to quantitatively assess closeness between the Albanian varieties based on the loanwords attested in them and to interpret the findings in the light of the "contact history" of Albanian. To this end, we survey the loanwords of various origin, i.e. Balkan Slavic, (Medieval and Modern) Greek, (Ottoman) Turkish, and (Western and Balkan) Romance, found in the Dialectological Atlas of the Albanian Language (DAAL) [Gjinari et al. 2008].

In the next section, we give an overview of Albanian dialects and of the diachronic layers of loanwords in modern Albanian. *Section 3* describes the data extracted from the Atlas and discusses several methodological and technical solutions for their processing. *Section 4* describes the quantitative analysis methods used in our study. In *Section 5*, we present and discuss the results obtained for each of the aforementioned sets of loanwords by origin. The last section gives a summary of our findings and proposes directions for further research.

2. Background

2.1. The traditional classification of Albanian dialects

The Albanian language area falls into two large dialectal zones, the Northern zone, or Gheg (Alb. *gegë* 'of or pertaining to Gegëria, an ethnographic region encompassing central and northern Albania, or its inhabitants'), and the Southern zone, or Tosk (Alb. *toskë* 'of or pertaining to Toskëria, an ethnographic region including southern Albania, or its inhabitants'). The dialectal varieties spoken in Northern Albania, Kosovo, Montenegro, and Southern Serbia, as well as the majority of Albanian varieties of North Macedonia belong to the Gheg zone, while the varieties of Southern Albania, Greece, and the Ohrid-Prespa area in North Macedonia are Tosk [1].

[1] The Albanian language also has several historical diaspora varieties. The variety of Zadar (Croatia) belongs to the Gheg dialect. The Arbëresh variety of Italy, the variety

The Shkumbin river that crosses Albania from east to west forms the dividing line between the two zones, with Gheg spoken north and Tosk, together with a narrow strip of the so-called transitional varieties, south of the river.

Both Gheg and Tosk comprise several subdialects shown in *Figure 1* from [Rusakov 2013: 165], based on [Gjinari et al. 2007: 56, Map C] and [Gjinari, Shkurtaj 2000: 185]. Gheg includes Northern Gheg, which is further subdivided into western and eastern subdialects, Central Gheg, and Southern Gheg (called "Central Albania Gheg" in most Albanian dialect

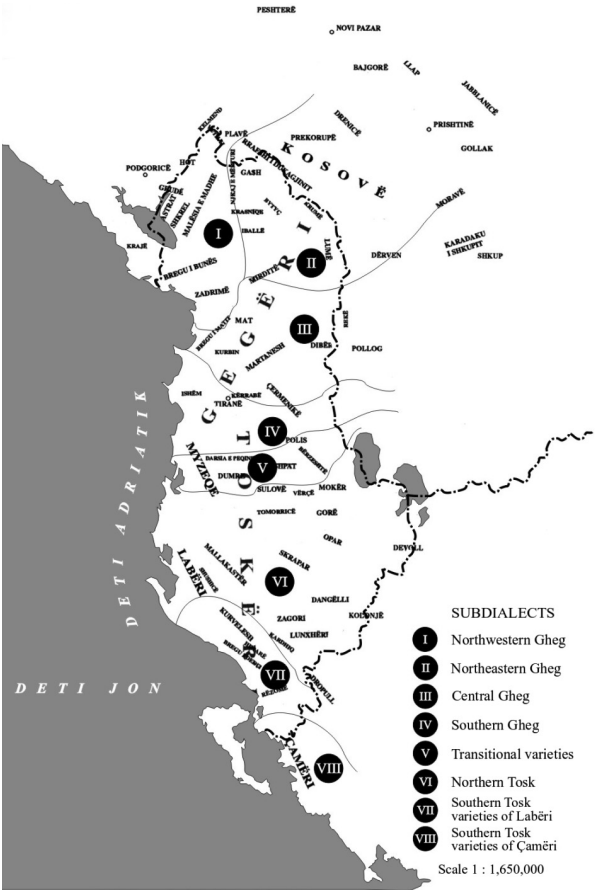

Figure 1. Classification of Albanian dialects

of the village of Mandritsa (Bulgaria), and the variety of Albanian spoken in Ukraine pertain to the Tosk dialect.

descriptions; see, for example, [Gjinari, Shkurtaj 2000: 186]). Tosk is divided into Northern Tosk with western and eastern subdialects, and Southern Tosk including Labëria and Çamëria subdialects.

The modern Gheg and Tosk dialects show mostly phonological and, to a lesser extent, morphosyntactic distinctions (see the full list in [Desnitskaya 1968a: 39–45; Gjinari, Shkurtaj 2000: 176–179]) that emerged at various stages of the Albanian language history. Though the initial dialectal split between Gheg and Tosk developed in the 8th–10th centuries, most features forming the present-day Albanian dialect landscape pertain to the first three centuries following the Ottoman invasion, i.e. the 16th–18th centuries [Gjinari, Shkurtaj 2000: 170–174; Rusakov 2013: 164]. Notably, the Tosk dialect is generally more balkanized than the Gheg dialect, probably as a result of the intensive multilingual contacts in the area south of the lakes Prespa and Ohrid where the local Tosk varieties of Albanian interacted with Greek, Macedonian, Arumanian, and Romani dialects throughout the second millennium AD [Lindstedt 2000: 234].

2.2. Loanwords in Albanian

Albanian belongs to languages with a large amount of lexical borrowings representing several chronological layers (see a detailed overview in [Demiraj 2013]). The earliest layer comprises words borrowed from **Ancient Greek** and **Latin**. There are slightly more than 30 Ancient Greek borrowings in Albanian (for example, *tym* 'smoke' < Gr. *thymos*) [Ölberg 1972], while the Latin borrowings are very substantial in number. According to various lists, the number of Latin etymons in Albanian shall be no less than 600 [Mihăescu 1966; Haarmann 1972; Landi 1989; Vătăşescu 1997; Bonnet 1998]. Latin loanwords entered the Albanian lexicon within the period of intensive Albanian-Latin language contacts, which might have started at the beginning of the 1st century AD, after the final incorporation of the Western Balkans into the Roman state, and lasted until the 5th–6th centuries AD, eventually taking the form of Albanian-(Proto)Rumanian contacts [Rusakov 2017: 125]. The Latin borrowings in Albanian penetrated into virtually all semantic fields, and most part of these borrowings can be found in all Albanian dialectal varieties. Cf. *gjyq* 'trial, court' < Lat. *iūdicum*, *mjek* 'physician' < Lat. *medicus*, *vij* 'come' < Lat. *venio*, and many others.

Slavic loanwords in Albanian, which have been the object of numerous studies such as [Selishchev 1931; Jokl 1934; Desnitskaya 1968b; Svane

1992; Ylli 1997; Sobolev 2012] and many others, are a result of language contacts between Albanian and South Slavic that, having started after the Slavic expansion into the Balkans in the 6th and 7th centuries, in a sense continue until now. Historical-phonetic criteria only make it possible to determine the time of borrowing for a small group (ca. twenty) of Slavic words that entered Albanian before the 10th–11th centuries, whereas the time of borrowing remains unclear for the majority of Slavisms. Unlike Latin borrowings, many Slavic loanwords show a clear dialectal distribution. Xhelal Ylli [1997] suggests that only a quarter of some 1000 Slavic borrowings are spread among all or almost all Albanian dialectal varieties (e.g. *oborr* 'yard', cf. Bg./Mc. and Srb. *обор*; *zakon* 'custom', cf. Srb. *закон*). The main source of Slavic loanwords for the southern Albanian varieties were Eastern South Slavic (Macedonian-Bulgarian) dialects, and for the northern varieties, Western South Slavic (Serbo-Croatian) dialects.

Middle Greek borrowings first penetrated Albanian in the early medieval period. Although they are attested mostly in the southern varieties of Albanian, some common Albanian words of Greek origin, such as *trëndafil* 'rose' < Gr. *triandafyllo*, are found in the north as well. Though the process of borrowing still continues in modern times through the ongoing contacts of Albanian with **Modern Greek**, such loanwords as, e.g. *fole* 'nest' < Gr. *folia*, occur almost exclusively in the Tosk varieties of Southern Albania and Northern Greece [Demiraj 2013: 166].

Italian loanwords, e.g. *barkë* 'boat' < It. *barca*, date back to the beginning of active contacts between Italian states and the coastal Albanian territories in the 11th century. Many of the early Italian borrowings came from the Venetian dialect (on Italian borrowings, see [Helbig 1903]). The few **Arumanian** borrowings attested in Albanian (for example, *milor* 'lamb' < Arum. *milior*) have mostly dialectal distribution. They result from contacts between Albanians and Arumanians that took place in both urban settlements in the Central and Southern Albania such as Elbasan and Voskopoja, and the surrounding rural areas, where people from the two ethnic groups would drive their livestock between grazing pastures.

After the Ottoman invasion of Albania in the late 14th — early 15th centuries, Albanian took a great influx of loanwords from Ottoman Turkish (words of Turkic origin, as well as Arabic and Persian borrowings in Turkish). **Turkish** loanwords in Albanian belong to various semantic fields and include terms related to economy, administrative activities, social and spiritual life, interjections and discourse markers, as well as some basic vocabulary words [Boretzky 1975, 1976; Dizdari 2005]. Cf. *bajrak* 'banner,

an administrative district in Albanian mountains' < Tr. *bayrak*, *borxh* 'debt' < Tr. *borç*, *sevda* 'love' < Tr. *sevda*, and many others.

3. Data: the source and processing

3.1. Data used in the study

The data used in the study come from the Dialectological Atlas of the Albanian Language, or DAAL [Gjinari et al. 2007, 2008], a two-volume atlas with 131 locations, or points (villages and some towns), in the Albanian language area and 14 points in the historical diaspora (Pešter in Serbia, Zadar in Croatia, Peloponnesus and islands in Greece, and Italy)[2]. The first volume [Gjinari et al. 2007] contains phonological and grammatical data collected in the 1970–1980s using a questionnaire with 65 questions on phonology and 80 questions on morphology and syntax [Idem: 437–453]. The second volume [Gjinari et al. 2008] maps the local terms for 260 lexical items belonging to 19 semantic fields (astronomic and meteorological terms, names of trees and plants, wild and domestic animals, household, kinship, body-part terms, names of material culture objects, etc.).

In our study, we focused on the Albanian varieties of the main area and chose 131 points, 93 of which are located in the Republic of Albania and in the adjacent part of Greece (Çamëria), 25 in Kosovo and in the south-western part of the Republic of Serbia (Preševo), seven in the Republic of North Macedonia, and six in the Republic of Montenegro. Diaspora varieties were not taken into consideration in the study.

For the subsequent analysis, we selected 218 of the 260 lexical maps from the second volume of the Atlas [Gjinari et al. 2008]. Excluded from further analysis were (1) maps capturing no lexical differences between the varieties as, e.g., map 412 showing a common Albanian lexeme *dhi* 'goat' [Gjinari et al. 2008: 112]; (2) maps with significant part of data missing, such as map 580 *bodec* 'metal tip of a goad' [Idem: 448]; and (3) maps with the predominance

[2] One speaker was interviewed for each point except point 17, the town of Shkodra, where two speakers were interviewed; the existing dialectal descriptions were used for points 24 (Vushtrri / Vučitrn) and 56 (Preševo). For some points situated in Greece and in the former Yugoslavia, speakers who had earlier migrated to the Republic of Albania were interviewed [Gjinari et al. 2007: 22].

of periphrastic descriptions, e.g. map 457 for ‘pregnant woman’, which, depending on the variety, is called *grua me bar* (woman.NOM.SG.INDF with burden.ACC.SG.INDF) ‘woman with burden’, *grua e ngarkuar* (woman.NOM.SG.INDF F.NOM.SG loaded.F.SG) ‘loaded woman’, *grua e lig* (woman.NOM.SG.INDF F.NOM.SG evil.F.SG) ‘sick (lit. “evil”) woman’, or *grue e rãn* (woman.NOM.SG.INDF F.NOM.SG heavy.F.SG) ‘heavy woman’ [Idem: 202].

3.2. Data processing

At the next step, we entered data from the selected 218 maps into an Excel table where each of the 131 dialectal varieties (rows) was characterized by 218 lexical features (columns). Each feature was represented by a word and its etymology (see a fragment in *Table 1* below).

Our study focused on the very fact of borrowing rather than on the subsequent development of the borrowed lexeme in the recipient variety. In the framework of this approach, we only considered etymologically distinct lexemes as different items and did not distinguish between phonetic variants and derivatives attested in the varieties. For example, the phonetic variants *mraul*, *mragəl*, and *muravəl* ‘ant’ [Gjinari et al. 2008: 182–183] were treated as one lexeme *mraul* in our table. In the same way, we treated the word *kłap* ‘trap, snare’ and its derivative *kłapei̯t͡skə* [Idem: 542–543], both referred to as *kllape* in the table. The elimination of phonetic variants is methodologically legitimate, given that in the majority of cases phonetic differences emerge by force of general phonetic processes independent of the words where they manifest (for example, a change of a single phonological feature may account for differences in a multitude of words). The situation is more complicated for derived words. On the one hand, where a loanword exhibits a certain derivational pattern in some varieties but not in others, this may reflect the degree of closeness between these groups of varieties. On the other hand, the choice of a derivational pattern may have more or less random character in any variety (“option selection”, according to [Matras 2005]), and in this case the occurrence of similar or different derived words in several varieties may provide no essential information about their degree of closeness.

Along with phonetic variants and derivatives, in the course of data processing we picked out (most of) periphrastic descriptions registered in the DAAL maps and excluded them from further analysis. For example, for map 483 *thjeshtër* ‘stepchild (stepson, stepdaughter)’, we had to mark as NA

("not available") all points where this notion was described as follows: *djalë i burrit* / *çun i burrit* (boy.NOM.SG.INDF M.NOM.SG.INDF husband:GEN.SG.DEF) 'son of husband', *gocë e burrit* (girl.NOM.SG.INDF F.NOM.SG.INDF husband:GEN.SG.DEF) 'daughter of husband', *fëmijë i burrit* (child.NOM.SG.INDF M.NOM.SG.INDF husband:GEN.SG.DEF) 'child of husband' [Gjinari et al. 2008: 254–255].

3.3. Etymologization

After the selection of maps and lexemes, we established the etymologies of words attested in the varieties using etymological dictionaries of Albanian [Çabej 1976, 1996, 2002, 2006, 2014, 1987/2017; Orel 1998; Topalli 2017], as well as dictionaries and monographs on loanwords in Albanian [Papahagi 1963; Svane 1992; Ylli 1997; Domosiletskaya 2002; Dizdari 2005]. A fragment of the data set is shown in *Table 1* (p. 285). All lexemes were divided into "inherited words", i. e., words of Albanian origin and old loanwords from Ancient Greek and Latin (labeled "alb")[3], and "borrowings". The borrowings were labeled by their origin as either Balkan Slavic ("slav"), Medieval and Modern Greek ("greek"), Ottoman Turkish ("turk"), or Romance. The words borrowed from Romance languages were split into two groups, those from Eastern Romance (from Arumanian, "arum") and from Western Romance (from Italian, Venetian, Dalmatian, etc., all referred to as "rom"). Several words, such as *kokomone* 'potato' in point 89 (Reka e Dibrës) or *zigur* 'young ram' in several points, were marked as NA, because we could not establish their etymologies.

In some cases, it was not possible to determine the immediate etymological source of a word. For instance, *flojer*(*e*) (*e këmbës*) 'shinbone', attested in some Albanian dialectal varieties, may be borrowed from either Greek or Arumanian. In such cases we had to make a decision on our own, sometimes in a more or less arbitrary way. In particular, *flojer*(*e*) was labeled as "greek" based on the occurrence of this lexeme only in the Albanian varieties spoken in Northern Greece; see [Gjinari et al. 2007: 298–299].

In a few cases we had to differentiate two words represented as one etymon in etymological dictionaries, such as *capë* and *çapë* 'hoe' [Gjinari et al. 2007: 442–443]. According to [Çabej 1987/2017: 10], *capë* may be "a native lexeme blended with some homophone foreign word", namely the Venetian

[3] The distribution of Ancient Greek and Latin borrowings does not essentially differ from that of genuine "inherited" words.

Table 1. Etymologization and labeling of lexical borrowings in the Albanian dialectal varieties

Point	Subdialect	Country, region	Lexeme_410	Etymology_410
			'young ram'	
2	NEG[4]	Kosovo	*rrundzak*	slav
3	NWG	Montenegro	*sheleg*	turk
17	NWG	Shkodër (Albania)	*qengj*	alb
80	SG	Kavajë (Albania)	*milor*	arum
135	NT	Skrapar (Albania)	*zigur*	NA
139	ST	Zagori (Albania)	*milor* / *sheleg*	arum/turk

zapa or Italian *zappa* 'hoe', also borrowed in Western South Slavic as *capa*. The second lexeme, *çapë*, is connected by some etymologists with Bg./Mc. *чапа* '(small) hoe' but Çabej supposes that this word could have emerged in Albanian as a phonetic variant of *capë* through alternation of affricates *c* and *ç* [Idem: 87]. We believe *capë* and *çapë* to most probably have different origins, the former being a Romance loanword (labeled as "rom") and the latter a Slavism ("slav").

3.4. Some source data problems

Some of the problems we confronted in the analysis of borrowings involved those of data representation (and representativeness) in DAAL that seem to be common for data sources such as dialect atlases in general.

The first point to make here is that our analysis is based on the limited, though rich and diverse, material of DAAL rather than on the Albanian lexicon in general. It is well known that words from different semantic fields demonstrate different degrees of borrowability [Tadmor 2009: 64–65]. Besides, the semantic distribution of loanwords depends on the social circumstances of the contact. According to [Svane 1992], for example, the majority of Slavic loanwords in Albanian belong to the semantic fields "Material culture", "Plants", "Animals", "Environment", and "Human body". These borrowings reflect the character of the cultural interaction between the

[4] Hereafter, we use the following abbreviations for the main subdialects of Albanian: NEG — Northeastern Gheg, NWG — Northwestern Gheg, CG — Central Gheg, SG — Southern Gheg, NT — Northern Tosk, ST — Southern Tosk.

Albanians and the Slavs after the latter arrived to the Balkans and suggest that this interaction mostly took place in agrarian communities. On the other hand, the politically and culturally dominant languages such as Greek and Ottoman Turkish contributed more to the semantic fields "Material culture", "Urban culture", "Administration", "Religion", etc. [Desnitskaya 1987; Demiraj 2013]. As DAAL does not cover the latter three semantic fields, this inevitably affects the overall distribution of loanwords in the Atlas. In addition, the very classification of words into semantic fields used in DAAL has its weaknesses. In a lexicon, some semantic fields contain only closed or nearly closed sets of lexemes ("Kinship") or relatively few words ("Time"), while other semantic fields ("Agriculture", "Vegetation") may include hundreds of words. DAAL, however, tends to present all semantic fields in a roughly the same way and gives practically equal numbers of maps for fields such as "Kinship" and "Agriculture".

Another problem of DAAL (or any other dialect atlas) data involves representation of semantic shifts. For example, map 413 'male goat' [Gjinari et al. 2008: 114] shows a "unique" lexeme *përç* (from Bg./Mc. *прч*, *прчи*; Srb. *прч*, *прчити*) in point 132 (Devoll) unattested in the other points. In Standard Albanian, however, *përç* means 'uncastrated male goat' and, according to Xhelal Ylli's study of Slavic loanwords in Albanian [1997: 189], this lexeme exists in most Tosk and some Gheg varieties. Obviously, 'male goat' and 'uncastrated male goat' are semantically close and figure as a more general and a more specific concept, respectively, while DAAL only provides a single map for the more general concept.

Finally, a more technical problem we faced was that in some cases the same word appeared in two different maps. For example, *bretkosë*, standing for 'frog' in Standard Albanian and in most Albanian varieties (map 444 in DAAL [Gjinari et al. 2008: 176–177]), also stands for 'toad' in several points on map 445 [Idem: 178–179]. Although frogs and toads pertain to different species, the similarity of their appearance may lead speakers to either generalize one of the terms for a generic-species category (as, for example, *лягушка* and *жаба* in colloquial Russian, according to [Russo 2016]), or mix the terms so that they can substitute each another in some varieties (see *Table 2*, p. 287).

For such cases, we decided that the very presence of a particular word (*bretkosë*, *zhabë*, *thithëlopë*, etc.) in a given variety is more important for our study, while registering two different features in cases like 'frog' and 'toad' would produce largely artificial distinctions between varieties. Thus, we opted for merging such words (and maps), as shown in *Table 3*.

Table 2. Words for ‘frog’ and ‘toad’ in DAAL [Gjinari et al. 2008: 176–179]

Point	‘frog’		‘toad’	
72	*bretkosë*	alb	*bretkosë*	alb
89	*zhabë*	slav	*zhabë*	slav
95	*bretkosë*	alb	*zhabë*	slav
121	*bretkosë*	alb	*thithëlopë*	alb
137	*zhabë*	slav	*bretkosë*	alb

Table 3. Words for ‘frog / toad’ in the final data set

Point	‘frog / toad’	
72	*bretkosë*	alb
89	*zhabë*	slav
95	*bretkosë*/*zhabë*	alb/slav
121	*bretkosë*/*thithëlopë*	alb/alb
137	*zhabë*/*bretkosë*	slav/alb

Six pairs of maps were merged in this way[5] bringing our final data set to a total of 212 lexical features.

4. Methods of quantitative analysis

4.1. Lexical closeness of varieties: calculating the distances

As the first step of our quantitative analysis, we measured the differences between the 131 varieties documented in DAAL in terms of the sets of borrowed lexemes attested. These differences were expressed as distances between variety pairs. The data were processed using ExcelVBA and R [R Core Team 2019].

The general principle of distance estimation for a pair of varieties was the following. For each of the concepts mapped in the Atlas we checked whether the two varieties employ the same or different lexemes. We counted

[5] Maps 444/445 ‘frog / toad’, 450/451 ‘wasp/bumblebee’, 459/460 ‘suckling / small child’, 487/489 ‘niece / granddaughter’, 494/495 ‘bone / bones (in the grave)’, 536/537 ‘tongs / small scoop used to shovel cinders and ashes’.

the number of matches and mismatches in the attested lexemes and divided the number of mismatches by the sum of matches and mismatches. Let us consider the data given in *Table 4*, where the symbols ✖ and ✓ correspond to mismatches and matches, respectively. In the table, there are three cases where the varieties employ different lexemes (3 mismatches), one case where the same lexeme is attested (1 match), and one case of a partial match (with 0.5 added to both matches and mismatches). As the sum of matches and mismatches equals 5, these data yield the distance equalling $3.5 \div (1.5 + 3.5) = 3.5 \div 5 = 0.7$[6].

Table 4. Measuring the distances between Albanian varieties based on lexicon

Meaning	**Variety 1**	**Variety 2**	**(Mis)matches**
'grapevine'	*hardhi*	*pjergull*	✖
'corn'	*drith* / *kalamboq*	*drith* / *misër*	✓/ ✖
'spike'	*karabush*	*kalli*	✖
'bunch of grape'	*vesh*	*verige*	✖
'awns'	*hala*	*hala*	✓

In our previous study [Rusakov et al. 2018], we used this method to estimate the degree of closeness between Albanian varieties based on all the lexemes mapped in DAAL, irrespective of their origin. In this study, we only consider borrowings and subgroups of borrowings rather than the entire set of lexical items. Obviously, the subsets of meanings corresponding to borrowed lexemes or to borrowings of a specific group in a pair of dialects will usually coincide only partially. For instance, there may be meanings expressed by a borrowed lexeme in one variety and by a lexeme of Albanian origin, in the other.

[6] Measures based on the proportion or number of (mis)matches between the two sets under comparison are widely used for distance estimation. In particular, a similar procedure is commonly employed to calculate the so-called lexico-statictical percentages to assess the closeness of presumably related languages based on the data gathered by wordlists [Dyen et al. 1992]. In dialectometry, a measure of this kind was first used in the classical paper by Séguy [1971] and now it is often referred to as the "relative identity value", a term introduced by Goebl, see, e.g., [Goebl 1993]. More generally, this type of distance measure can be said to be based on the so-called simple matching coefficient, which is equal to the number of matches divided by the total number of the variables compared. The distance measure we use in this study can be calculated by subtracting the simple matching coefficient from 1.

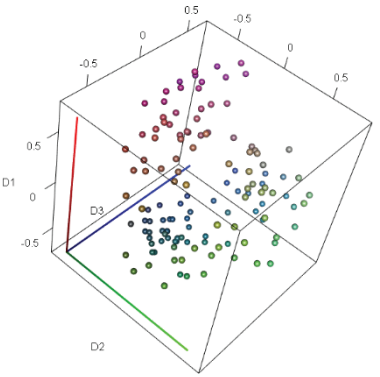

Figure 2. A three-dimensional MDS plot of distances
between Albanian varieties colored using the RGB-scheme

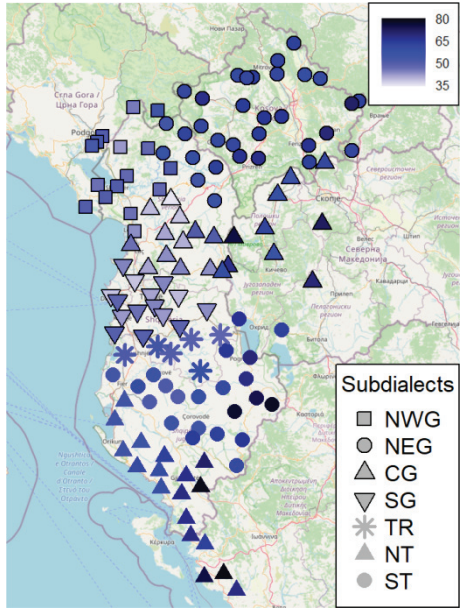

Figure 3. The overall number of loanwords in Albanian varieties

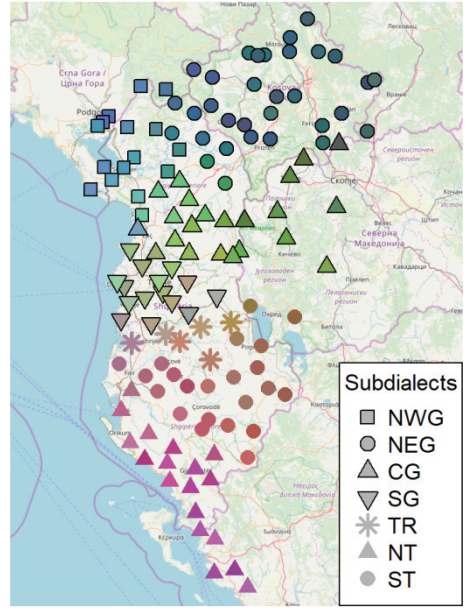

Figure 4. Closeness between Albanian varieties (based on the total sample of loanwords)

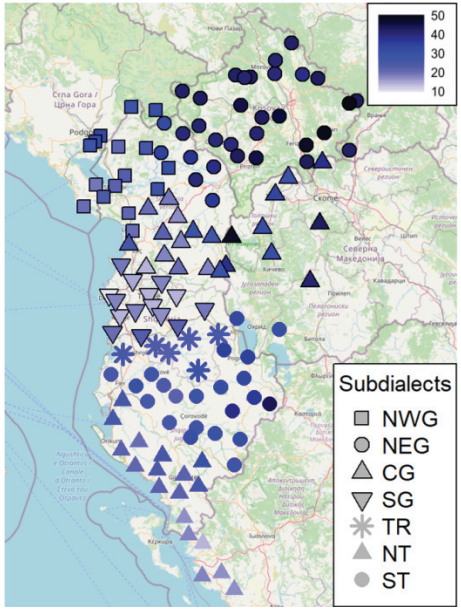

Figure 5. Slavic loanwords in Albanian varieties

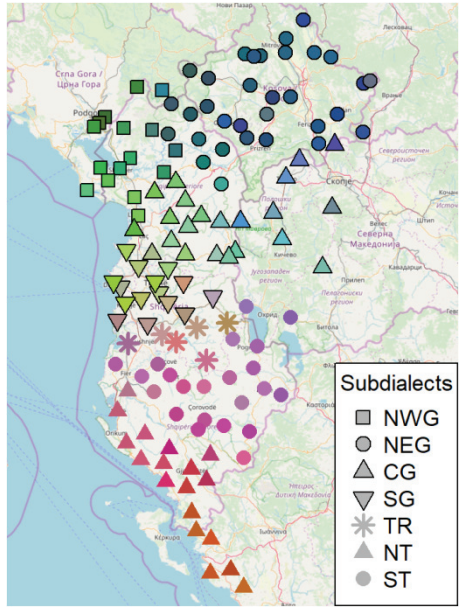

Figure 6. Closeness between Albanian varieties (based on Slavic loanwords)

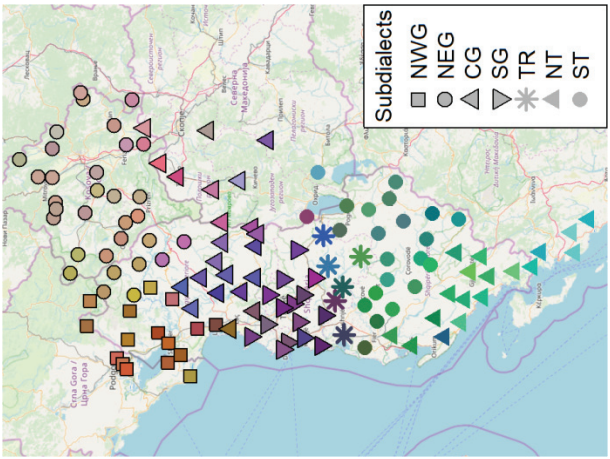

Figure 8. Closeness between varieties (based on Turkish loanwords)

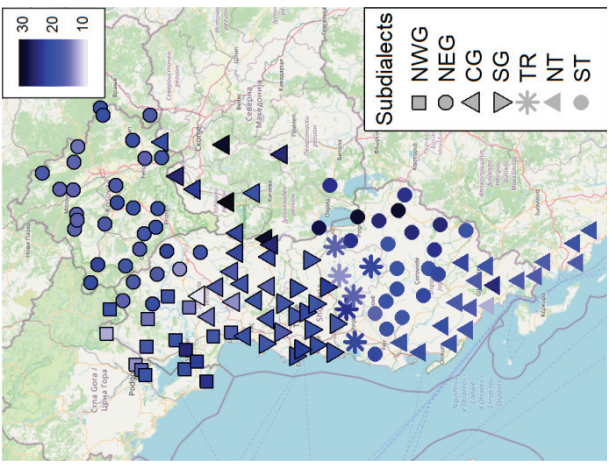

Figure 7. Turkish loanwords in Albanian varieties

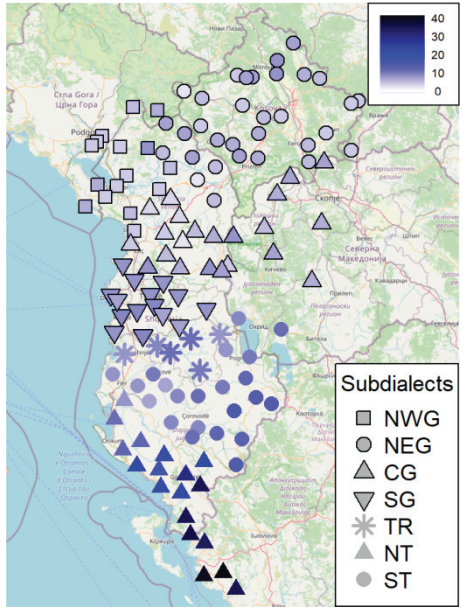

Figure 9. Greek loanwords in Albanian varieties: number of lexemes

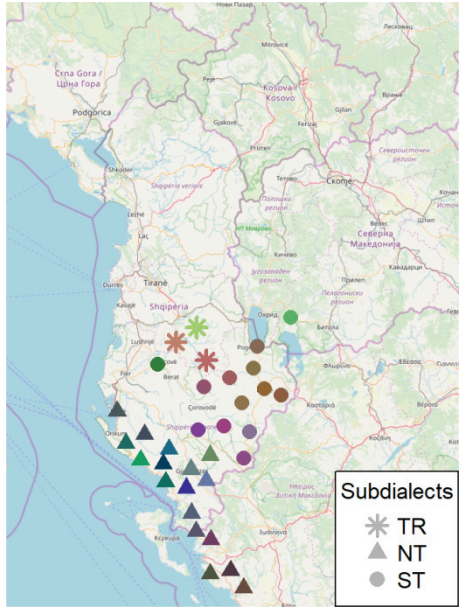

Figure 10. Closeness between Albanian varieties (based on Greek loanwords)

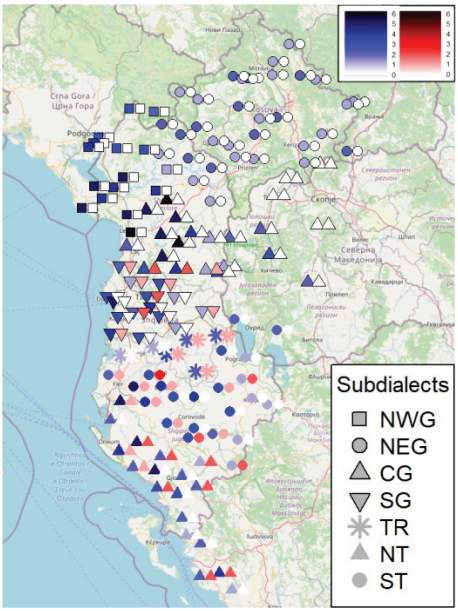

Figure 11. Romance loanwords in Albanian varieties

There are two possible approaches to calculating distances based on borrowings and their subgroups. The first is to consider only meanings expressed by a borrowed lexeme in both varieties and count matches and mismatches among borrowed lexemes only. The second approach is to take into account the meanings expressed by borrowings in either of the two varieties, i.e. also include the cases where one of the varieties uses a borrowing and the other, a lexeme of Albanian origin.

In essence, the two approaches address two different research questions. The first shows the differences between the dialects in terms of the set of borrowings (all borrowings or those of a specific origin) attested. The second reveals the differences both in the sets of borrowings (or their subgroups) and in the extent of their spread across the dialects. The latter approach inevitably results in larger distances between the dialects compared to the former, because the number of matching lexical items always remains the same while the number of mismatches is necessarily higher under the second approach.

In this study, we have chosen the second approach to distance estimation; given that distances of this type are calculated using a wider range of lexemes, they are more reliable, can be calculated for more variety pairs, and thus ensure a more comprehensive picture of lexical correspondences between the varieties. The result of such calculations is a distance matrix with pairwise distances between the points (dialectal varieties) in the data set. A fragment of the matrix is shown in *Table 5*.

Table 5. Distance matrix for the Albanian varieties in DAAL

	Point 1	**Point 2**	**Point 3**	**…**	**Point 131**
Point 1	0.0	0.5	0.3	…	0.6
Point 2	0.5	0.0	0.2	…	0.4
Point 3	0.3	0.2	0.0	…	0.1
…	…	…	…	…	…
Point 131	0.6	0.4	0.1	…	0.0

4.2. MDS plotting and geographic data mapping

At the second step, we visualized the distances between varieties for further qualitative interpretation. For visualization, we used the multidimensional scaling (MDS) algorithm that allows plotting the distances between

object pairs in two- or three-dimensional space aiming at minimal distortion of the original distances (see [Nerbonne, Wieling 2018] on the application of MDS in dialectometry). We performed the MDS-analysis and visualized its results using the packages *smacof* [de Leeuw, Mair 2009] and *rgl* [Adler, Murdoch et al. 2018] for R. We used the ratio MDS, which is the default option for the mds() function in *smacof*; in MDS models of this type, the ratios of the original distances "correspond to the ratios of the distances in the MDS space" [Mair et al. 2015: 4].

We used the MDS algorithm for computing coordinates of the varieties to plot them in a three-dimensional space. We opted for a three-dimensional solution, as it affords rendering the original distances with more precision compared to the two-dimensional solution while allowing for visual representation. The resulting MDS plot for distances between the varieties based on all the borrowings is shown in *Figure 2* (see the color insert). The colors of the points correspond to their position on the plot rendered using the RGB-scheme, with the red, green, and blue colors corresponding to the first, second, and third dimensions, respectively; see [Nerbonne et al. 1999] for a similar visualization method. The closer are the points on the plot, the more similar are their colors. As the next step, using the packages *rosm* and *prettymapr* [Dunnington 2017a; 2017b] for R, we plotted the points on the geographic map using the colors that correspond to their closeness on the MDS plot, see *5.1*. This way, we can assess the relation between the geographical proximity of the varieties and their similarity in terms of borrowings.

Four geographic maps were created using this technique: a general map of lexical borrowings and maps for the three groups of borrowings based on their etymology, i. e. Slavic, Turkish, and Modern Greek loanwords (see *Figures 4*, *6*, *8*, and *10* in the color insert). The quality of MDS-models is usually assessed by the stress value that corresponds to the degree of the original distances distortion with respect to the distances plotted. The stress values for the models used to plot the maps below equal 0.15 for all and for Slavic borrowings, 0.16 for Turkish, and 0.19 for Greek borrowings. With the traditionally accepted borderline stress value of 0.20, the above values indicate that the correspondence between the original and the plotted distances is tolerable but far from perfect, while the results for Greek borrowings should be interpreted with special caution.

Our maps use different symbols for Albanian subdialects, defined in *2.1* above in accordance with the traditional dialectological descriptions [Desnitskaya 1968; Gjinari 1989; Gjinari, Shkurtaj 2000] (in particular,

symbols with borders were chosen for Gheg and without borders, for Tosk). Thus, these maps allow analyzing the relationships between the three aspects of Albanian dialects' distribution: geographic proximity, affiliation in terms of dialect classification, and linguistic closeness in terms of the sets of borrowings [7]. Still, at this stage, our generalizations will only be of a preliminary and exploratory nature, with their statistical validity to be tested in the future.

In addition to the quantitative analysis performed in R, we calculated the absolute number of borrowings (i.e., all borrowed lexemes and loanwords from different etymological groups) to produce five maps showing the overall number of loanwords and the numbers of borrowings from Slavic, Turkish, Greek, and the Romance languages (see *Figures 3*, *5*, *7*, *9*, and *11* in the color insert). This makes it possible to identify the varieties, more and less subject to lexical borrowing, and visualize the spread of loanwords from different languages over the Albanian-speaking area. The numbers of borrowings are shown with gradations of blue, and the same symbols as in the MDS maps were used to distinguish between the subdialects.

5. Loanwords in Albanian varieties: analysis and results

5.1. The overall number of loanwords and the closeness between Albanian varieties

This subsection considers the numbers and distribution of lexical borrowings across Albanian varieties irrespective of their origin to assess the degree of closeness between the varieties by means of the quantitative analysis described in *4.1*. The number of borrowings in each variety is shown in *Figure 3* (see the color insert) as a percentage of the total number of 212 words, or lexical features, while the results of the quantitative study are plotted on the geographical map in *Figure 4* (see the color insert).

The total number of loanwords in our sample of 212 words varies from 38 to 83. There are no varieties lacking or showing very small numbers

[7] Another method often used to visualize and explore distances is hierarchical agglomerative clustering, see the discussion of this method from the perspective of dialectometry in [Nerbonne, Wieling 2018: 401–402]. This method facilitates analysis of linguistic closeness against dialect classification but does not allow for a straightforward comparison of these two facets with the spatial proximity.

of loanwords, which agrees well with the general knowledge about the rich contact history of Albanian as a whole.

A more or less significant increase in the number of loanwords is observed in the peripheral varieties spoken in areas with ongoing or recent language contact. These include the Central Gheg varieties in North Macedonia (especially those on the Albanian-Macedonian state border), the Northeastern Gheg varieties spoken in Kosovo and the neighboring parts of Albania, the Northwestern Gheg of Montenegro, the Southern Tosk varieties of Çamëria and to some extent Labëria, and the Northern Tosk varieties of the regions of Korça and Devoll in the southeast of Albania. An opposite picture is characteristic of the Northwestern Gheg in Albania, inland and coastal Northwestern Tosk varieties, and especially of the Central Gheg and Southern Gheg varieties spoken in Albania that show a moderate number of lexical borrowings.

Our calculation of the median number of borrowings in each dialectal subgroup reflects this center vs. periphery distinction even more vividly. Thus, the median number of borrowings in the Central Gheg varieties in Albania is 46 against 65 for those in North Macedonia. At the same time, the median value distribution shows considerable differences in the number of loanwords between the varieties spoken in Albanian areas remote from the actual contact zones. These differences are correlative with the traditional dialect classification, mostly based on the phonetic and grammatical isoglosses. In particular, the median numbers of borrowings in Northern Tosk varieties are 68.5 for the Northeastern and 56 for the Northwestern Tosk area. A similar, though less pronounced, difference is observed in Northern Gheg: Northeastern Gheg varieties (both in and outside Albania) show higher median numbers of borrowings than Northwestern varieties (Northeastern Gheg: 66 in Kosovo and 60 in Albania; Northwestern Gheg: 54 in Montenegro and 50 in Albania). These differences, presumably, may be traced to a relatively deep chronological level, namely to the time when the modern Albanian dialect landscape was being formed. One of the very likely factors in the process may have been the (lack of) contacts with other languages.

Our comparison of the median values for the number of borrowings in Gheg and Tosk shows these two dialect zones differing in the number of loanwords. While most of the ongoing contact areas (Kosovo, the larger part of North Macedonia, and Montenegro) lie in the Gheg zone, the Tosk zone attests more loanwords (the difference between the Gheg and Tosk varieties is statistically significant; Wilcoxon-Mann-Whitney Test, $W = 857$,

$p < 0.001$): the median for the Gheg zone is 56 against 66 for the Tosk zone. On the one hand, this difference reflects the presence of several Gheg zone areas with relatively low numbers of loanwords (specifically, m=46 for the Central Gheg spoken in Albania and 47, for the Southern Gheg). On the other hand, it supports the idea of more intensive language contacts in the history of the Tosk Albanian; see *2.1* on a stronger balkanization of Tosk and [Rusakov et al. 2018] on the contact-induced grammatical and phonetic simplicity of Tosk varieties.

An interesting question that arises in connection with the contact history of Albanian is whether the amount of loanwords in Albanian varieties or groups thereof reflects the type of bilingual situation in the corresponding areas. Compare, for example, a relatively low count of loanwords in Montenegro (m=54) against much higher rates in Çamëria (m=71). In the south of Montenegro, situations of balanced Slavic-Albanian bilingualism, with a relative equality of both languages, could have existed in the recent past and can still be found in some rural communities [Morozova, Rusakov 2018], while the situation in the Greek part of Çamëria, on the contrary, has always been unbalanced with a strong Greek dominance.

The map in *Figure 4* (see the color insert), based on the results of the three-dimensional MDS analysis (see *Figure 2* in the color insert), provides a closer look at the traditionally defined Albanian subdialects as regards their homogeneity or diversity.

Figure 4 shows the Northeastern Gheg subdialect as a very homogenous group of varieties. The points marked by closely matching colors represent varieties with (almost) equal numbers and similar sets of lexical borrowings. A part of the explanation may lie in the common contact history of the Northeastern Gheg zone involving a strong influence of Serbian dialects (see *5.2*) that resulted in a common borrowed vocabulary in the different local varieties of Albanian.

The Central Gheg subdialect falls into two parts as was already observed in *Figure 3* and shown by our median-value estimation of the number of borrowings in Albanian subdialects. The eastern part covers North Macedonia with its continuous Slavic-Albanian contacts and the western part, isolated inland zones in Albania. A tentative grouping including some Northwestern and Central Gheg varieties can be observed close to the seacoast. The Southern Gheg subdialect shows less homogeneity than any other group.

The Tosk dialect zone reveals certain differences between the western (especially coastal) and eastern parts of the Northern Tosk. The Southern Tosk subdialect zone shows distinctions between the varieties spoken in the

northern and the southern parts of the Labëria area (see *Figure 1*), close to the neighboring Northern Tosk and Çamëria varieties, respectively.

In general, both the distribution of the numbers of borrowings and the closeness of the varieties support and further specify rather than contradict the traditional Albanian dialect classification. Our data confirm the majority of the established dialect subgroups, often showing the gradual nature of the borderline varieties and in some cases suggest additional areal distinctions within the traditional subgroups.

5.2. Slavic loanwords

The number of Slavic loanwords, as seen from *Figure 5* (see the color insert), varied from 10 (in the town of Durrës, Southern Gheg) to 50 (in the Northeastern Gheg variety of Hogosht in Kosovo). The medians calculated for the main dialect areas ranged from 17 (Çamëri) to 43 (Kosovo).

Figure 4 (see the color insert) shows the smallest number of Slavisms in areas with the modest loanword numbers overall[8] (Northwestern and Central Gheg, the western part of Northern Tosk), as well as in contact areas with languages or dialects other than Slavic (Labëria and especially Çamëria in Southern Albania and Northern Greece, see *5.4*). This indicates that Slavic loanwords' distribution is determined by an areal factor, i.e. by the intensity of contact between Albanians and Slavs who inhabit(ed) certain parts of the main geographic area of Albanian (as distinct from the distribution of Turkish borrowings that do not cluster geographically; see *4.3*).

It is also of note that the amount of Slavic lexical borrowings in Albanian varieties of Montenegro is quite moderate, while the varieties of the other areas of ongoing Albanian-Slavic contact, such as Kosovo or North Macedonia, show the highest rates of Slavisms. This difference may point to the different types of contact situations in the eastern (Kosovo and North Macedonia) and western (Montenegro) parts of the historical Albanian-Slavic contact area, just as in the clearer case of distinctions between Montenegro and the Greek part of Çamëria, discussed in *5.1*.

Our quantitative analysis of closeness between Albanian dialect varieties based on Slavic loanwords (*Figure 6* in the color insert) produced

[8] This effect is in part explained by the fact that Slavic borrowings are more numerous than those of the other etymological groups, and their distribution largely determines the distribution of borrowings in general.

results largely similar to those obtained for borrowings in general; see *Figure 4* in the insert. Some differences can be only noticed in the southern and central parts of the Albanian geographic area. First, Labëria in the Southern Tosk zone is more distinct from Çamëria and appears to be closer to Northern Tosk varieties. This means that the inventory of Slavic borrowings in the Tosk varieties south of the present-day border of Albania and Greece, where the larger part of Çamëria lies, differs from that in the main part of the Tosk dialect zone, located in Albania. Second, a number of the Southern Gheg varieties share some Slavic loanwords with the neighboring Northern Tosk and transitional varieties and thus appear to be closer to them than in *Figure 4*. Finally, a more or less discrete "Montenegrin" group can be distinguished within the traditional Northwestern Gheg subdialect encompassing the varieties spoken in Montenegro and in the northwest of Albania.

In general, we may postulate four distinct areas of stronger Slavic influence on Albanian that include Montenegro (where it is less evident than in the other three areas), Kosovo, the northwestern part of North Macedonia, and the Southeastern Albania together with the southwestern part of North Macedonia (the Ohrid and Prespa lakes area where Tosk varieties are spoken). This distinction may be attributed to contacts with different Slavic dialects of the Eastern and Western South Slavic dialect continua (also see *2.2*) and to the different time depth of contact in particular areas. The Albanian varieties spoken in Montenegro and Kosovo owe a large part of their specific Slavic loanwords to contacts with speakers of various Serbo-Croatian dialects. The "Montenegrin" area crystallized as a result of the long-standing ethnic symbiosis of Albanian and Montenegrin tribes and more or less balanced Albanian-Slavic bilingualism [Morozova, Rusakov 2018]. In Kosovo, the Albanian-Slavic contacts began in the late medieval period and had an extremely complicated and diverse character. The source of the majority of Slavic loanwords in North Macedonia and Southern Albania are Bulgarian-Macedonian dialects. The presence of Albanian-speaking populations in the northwestern and western parts of the modern North Macedonia is mainly a result of the influx in the 18th–19th centuries [Selishchev 1931], although the earliest evidence of Albanian population in Macedonia is attested in medieval sources. Thus, large-scale Albanian-Slavic contacts in this area began late and must have been quite intensive. As for the southeast of Albania and the southwest of North Macedonia, this area belongs to the larger multilingual and multiethnic zone of intensive contact, located around the Ohrid and Prespa lakes (see *2.1*), and the mutual

influence of the local Albanian and Macedonian varieties, both in structure and lexicon, must have lasted here for some centuries.

In sum, the distribution of Slavic borrowings in terms of their number and the closeness of the varieties shows a strong areal pattern though, in contrast to borrowings in general, we observe several groupings of varieties cutting across the traditional dialectal divisions, in particular among the Central Gheg, Southern Gheg, and Southern Tosk subdialects.

5.3. Ottoman Turkish loanwords

The quantitative distribution of Ottoman Turkish borrowings across Albanian dialect varieties is shown in *Figure 7* (see the color insert). The number of Turkish loanwords varies from 10 in a Central Gheg point in Albania (Gojani i Epërm in Mirdita) to 27 in a Central Gheg variety in North Macedonia (the village of Ravenë in Pollog). The medians in the majority of the dialect groups vary from 16 to 18 with the exception of Central Gheg varieties in North Macedonia (m=23) and Northeastern Tosk varieties (m=22.5), which makes the distribution of Turkish loanwords across the subdialects more even than that of Slavic borrowings.

Figure 7 reveals no clear areal distribution of Turkish borrowings across Albanian varieties (in contrast to the Slavic borrowings in 5.2). More intensive borrowing is registered in the strongly balkanized Northeastern Tosk varieties spoken around and south of the lakes Ohrid and Prespa, as well as in the rural areas of North Macedonia and villages located in the borderline Central Gheg regions of Albania, Drimkoll and Golloborda. Modern towns such as Debar, Shkodra, Ulcinj, Korça, Pogradec, and Delvina, old Ottoman economic and cultural centers, also show relatively high rates of Turkish borrowing. This suggests that the degree of Turkish influence on the lexicon of Albanian varieties stems from historical and cultural rather than areal factors.

Our MDS analysis shows the closeness between varieties to roughly correspond to the traditional Albanian dialect classification. Notably, homogeneous (to various degree) clusters in *Figure 8* (see the color insert) include the varieties belonging to different dialect subgroups. One of such clusters comprises Kosovo varieties pertaining to Northeastern Gheg subdialect and several Central Gheg varieties spoken in the northern part of North Macedonia. Another group includes Tosk varieties without a clear differentiation between Northern and Southern Tosk and with a slight deviation of Çamëria

in the south from the rest of the area. With a rather homogeneous group of varieties in the western part of the Central Gheg area, the far western Central Gheg points seem to be closer to Northwestern Gheg. Lastly, a rather homogeneous group of the Northwestern Gheg varieties are spoken around the lake Skadar (in Montenegro and in the Shkodra area in Albania), while the rest of the Northwestern Gheg varieties show more similarity with the neighboring Northeastern Gheg area. Southern Gheg subgroups demonstrate several small groupings marked by the same color that nonetheless do not form a homogenous cluster. The transitional varieties and most Gheg and Tosk varieties spoken in North Macedonia are fairly diverse and sometimes stand apart from all their nearest neighbors.

The distribution of Turkish loanwords needs further analysis. We can only speak here of at least two big areas of Turkish influence: one in Kosovo, Northeastern Albania, and in the adjacent parts of North Macedonia, and the other in the Tosk zone. The lack of areality in the distribution of Turkish loanwords may be due to the fact that all these words entered Albanian varieties very late and their lexical "competition" with native words and alternative Ancient Greek, Latin, Balkan Slavic, Medieval Greek, and Western and Eastern Romance borrowings had different outcomes in different regions.

5.4. Medieval and Modern Greek loanwords

As mentioned in *3.3*, we marked a few Ancient Greek words attested in Albanian varieties as "native" ("alb"). Therefore, the maps below show only those lexemes of Greek origin that entered Albanian as of the Middle Ages and later.

The majority of Greek loanwords in Albanian expectedly belongs to Southern Tosk varieties that remain in contact with Greek. As seen in *Figure 9* (see the color insert), several distinct groups of Albanian varieties can be arranged in the following hierarchy in the order of the descending numbers of Greek borrowings: Çamëria > Labëria > Northern Tosk and transitional varieties > Southern Gheg > other Gheg varieties. The number of Greek loanwords visibly decreases in proportion to the spatial distancing from the Albanian-Greek border and the southern Albanian regions with Greek-speaking population.

Only the Southern Tosk (Çamëria and Labëria), three Northeastern Tosk and two transitional varieties have 10 or more Greek loanwords each. The largest number of Greek borrowings in Labëria is 35 (Nepravishtë

in the regjion of Gjirokastra, and Pandelejmon in the region of Saranda), while the maximum rate in Çamëria is 40 (the village of Karbunarë on the Albanian-Greek border). The median values for Labëria and Çamëria are 22 and 36, respectively. By contrast, the median value for all Gheg varieties is 6, and the maximal number of Greek borrowings (9) is found around urban centers such as Tirana, Elbasan, and Kavaja (all pertaining to Southern Gheg).

For our MDS analysis, we chose only those varieties where the number of Greek loanwords was equal to or more than 10[9]. As mentioned above, all these varieties belong to the Southern and Northern Tosk subdialects or to the group of transitional varieties. *Figure 10* (see the color insert) shows several vague groupings such as Çamëria, the northernmost part of Labëria, and the Pogradec — Korça — Devoll zone in the southeast of Albania. The results are rather preliminary due to the limited material analyzed.

5.5. Romance loanwords

As the number of lexical borrowings from Western and Eastern (or Balkan) Romance languages was very low in our sample, these data were insufficient for a quantitative analysis of closeness between Albanian varieties. However, the distribution of these borrowings across the Albanian geographic area (*Figure 11*, see the color insert) reveals some interesting features to be verified in the future based on a more representative data set.

The Balkan Romance (Arumanian) borrowings are insignificant in number (max=3) and are found in the Southern and Central Albania where a few representatives of the Arumanian minority still reside. They are not attested in most Northern Gheg and Central Gheg varieties (except in the three Central Gheg points situated along the "border" with the Southern Gheg subdialectal area), despite the fact that the Arumanian-speaking population existed in Kosovo and Montenegro in the Middle Ages and at the beginning of the modern times and still exists in North Macedonia.

Lexical borrowings from Western Romance languages can be found in almost all Albanian varieties. The majority of them show between 1 to 3

[9] This value was chosen as an arbitrary borderline to exclude the varieties where the number of Greek borrowings is too low to be analyzed and the distances may be less reliable. As *Figures 9* and *10* show, the analysis of this subset of borrowings was based only on varieties that are geographically closer to Greece.

borrowings each. A higher rate of Western Romance borrowings (4 to 6) is observed in a small group of Northwestern Gheg varieties spoken in Albania (the regions of Shkodra and Zadrima) and in the Southern Tosk varieties of Vlora and Mallakastra, i.e. in the seaside regions of Albania that hadcontacts with various Italian regions in the medieval times. Interestingly, a relatively high rate of Western Romance loanwords is attested in the most isolated parts of the Central Gheg zone such as Lura, Mat, and Mirdita. This fact may shed light on the origin of the Central Gheg group and probably points to their closer connection with the seaside parts of the Northern Albania in the period of their formation, which covers the Skanderbeg's time and the first centuries after the Ottoman invasion, according to [Beci 1965].

6. Conclusions

In this study, we have undertaken a quantitative analysis of the geographical distribution of borrowings in Albanian dialect varieties and proposed an interpretation of the results in the light of the contact history of Albanian. We focused on the borrowings in general, as well as on several subgroups of borrowings of different origins, i.e. Slavic, Turkish, Greek, and Romance borrowings. As our starting point, we took the traditional classification of Albanian dialects and the existing knowledge of the contact history of Albanian dialectal zones, see [Desnitskaya 1968; Gjinari 1989; Gjinari, Shkurtaj 2000; Rusakov 2013]. Using the Dialectological Atlas of the Albanian Language [Gijnari et al. 2008] as a data source, we analyzed the number of borrowings in dialect varieties and the degree of closeness between varieties in terms of the extent and sets of borrowings. The main findings of the study are as follows.

The quantitative distribution of borrowings shows a clear areal pattern where the periphery of the Albanian-speaking area is more prone to lexical borrowing than the center. Our data on the closeness between varieties based on the whole set of borrowings mostly coincides with the traditional dialect classification while adding a number of distinctions within the long-established groups. Thus, the Central Gheg varieties of North Macedonia appear to make a distinct group. Our data also suggest tentative subdivisions within the Northern and Southern Tosk subdialects in the Tosk dialect area.

The distribution of Slavic borrowings in terms of their number and the closeness between varieties largely coincides with the overall distribution

of borrowings and reveals a clear areal pattern. At the same time, several groupings of varieties cut across the traditional dialect classification, especially in the Central Gheg and in the Southern Tosk areas. Our analysis of closeness also discovered several well-distinguished zones of strong Slavic influence on Albanian, including Montenegro, Kosovo, North Macedonia, and the southeast of Albania.

Kosovo in general behaves as a very well-defined and closely-knit area, both from the point of view of the whole sample of borrowings and the different etymological groups. By contrast, the Southern Gheg zone is the least homogeneous area in the Albanian dialectal landscape. To a certain extent, it may be explained by the facts of the Albanian ethnic history. The Albanian-speaking communities of Kosovo, including those that arrived with the numerous waves of migration from the Northern Albania, adopted more or less identical sets of borrowings in their contacts with the homogeneous Slavic population of this territory. The Southern Gheg zone population largely also have a migrational background having arrived to these desolated lands after the Skanderbeg's wars. In this zone, however, they had no significant neighboring populations to interact with, and the contact history of the Southern Gheg subdialect throughout the Ottoman period was, in essence, limited to a rather superfluous influence of the high-prestige Turkish language.

In contrast to Slavic borrowings, (Ottoman) Turkish loanwords show no clear areal distribution. Their high concentration is observed in the Albanian varieties of North Macedonia and in several, mostly Southern Albanian, urban centers.

Greek loanwords are concentrated in the Southern Tosk dialect area (Labëri and Çamëri), though our analysis of closeness between these varieties does not show any clear groupings, probably due to the scarcity of the data.

The coastal Northwestern and Central Gheg varieties (jointly referred to as "Western Gheg" in [Gjinari 1989]) demonstrate more closeness to each other than to the other varieties of the corresponding subdialects, especially when it comes to the numbers of Western Romance borrowings. Relatively high numbers of Western Romance loanwords is also observed in the isolated Central Gheg varieties, which may throw light on the early history of these dialects. Arumanian borrowings can be found in the Central and Southern Albania.

An interesting empirical observation resulting from our analysis is that the closeness of the varieties based on the overall sample of borrowings corresponds to the traditional dialect classification to a higher degree than that

based on any of the specific subgroups of borrowings. As Slavic and Turkish borrowings are the two most numerous groups, the closeness based on the overall sample of borrowings may be a result of the superposition of the distributions observed for these two groups. The Slavic borrowings data often points to distinctions more fine-grained than the traditional dialectal division, whereas the distribution of Turkish borrowings in terms of closeness is too blurred for any clear distinctions to be detected. Therefore, the combination of these two distributions yields a picture that is in-between these two opposite effects and converges on the groupings that can be recognized as traditional dialectal subdivisions. More generally, this observation may suggest that while the distributions of specific groups of borrowings primarily reflect the particular contact scenarios, the cumulative effect of these distributions reveals variety groupings that share a common contact history, and it is these groups that are more likely to correspond to the traditional dialect groups defined on the basis of their grammatical and phonetic features.

List of abbreviations

Alb. — Albanian, Arum. — Arumanian, Bg. — Bulgarian, CG — Central Gheg, DEF — definite form, F — feminine, GEN — Genitive, Gr. — Greek, INDF — indefinite form, It. — Italian, Lat. — Latin, M — masculine, Mc. — Macedonian, NEG — Northeastern Gheg, NOM — Nominative, NT — Northern Tosk, NWG — Northwestern Gheg, SG — singular, SG — Southern Gheg, Srb. — Serbian, ST — Southern Tosk, Tr. — Turkish.

References

Adler, Murdoch et al. 2018 — D. Adler, D. Murdoch, O. Nenadic, S. Urbanek, M. Chen, A. Gebhardt, B. Bolker, G. Csardi, A. Strzelecki, A. Senger. rgl: 3D Visualization Using OpenGL. R package version 0.99.16. 2018. Available at: https://CRAN.R-project.org/package=rgl (accessed on 16.12.2019).

Beci 1965 — B. Beci. Mbi katër inovacione fonetike të të folmeve të Gegnisë së Mesme [On the four phonetic innovations in the varieties of the Central Gegni]. A. Kostallari (red.). *Konferenca e parë e studimeve albanologjike: Tiranë, 15–21 nëndor 1962* [First conference on Albanological studies, November 15–21, 1962]. Tiranë: Instituti i Historisë dhe i Gjuhësisë, 1965. F. 261–269.

Bonnet 1998 — G. Bonnet. Les mots latins de l'albanais. Paris: L'Harmattan, 1998.

Boretzky 1975 — N. Boretzky. Der türkische Einfluss auf das Albanische. Teil 1: Phonologie und Morphologie der albanischen Turzismen. (Albanische Forschungen 11). Wiesbaden: Otto Harrassowitz, 1975.

Boretzky 1976 — N. Boretzky. Der türkische Einfluss auf das Albanische. Teil 2: Wörterbuch der albanischen Turzismen. (Albanische Forschungen 12). Wiesbaden: Otto Harrassowitz, 1976.

Çabej 1976 — E. Çabej. Studime etimologjike në fushë të shqipes [Etymological studies in Albanian]. Bleu 2: A–B. Tiranë: Akademia e Shkencave e RP të Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, 1976.

Çabej 1996 — E. Çabej. Studime etimologjike në fushë të shqipes [Etymological studies in Albanian]. Bleu 4: DH–J. Përgat. për shtyp dhe pajisur me treguesit e fjalëve nga S. Mansaku dhe A. Omari. Tiranë: Akademia e Shkencave e R.Sh., Instituti i Gjuhësisë dhe i Letërsisë, 1996.

Çabej 2002 — E. Çabej. Studime etimologjike në fushë të shqipes [Etymological studies in Albanian]. Bleu 6: N–RR. Përgat. për botim S. Mansaku, A. Omari dhe B. Çabej. Tiranë: Akademia e Shkencave e R.Sh., Instituti i Gjuhësisë dhe i Letërsisë, 2002.

Çabej 2006 — E. Çabej. Studime etimologjike në fushë të shqipes [Etymological studies in Albanian]. Bleu 7: S–ZH. Përgat. për botim S. Mansaku, A. Omari dhe B. Çabej. Tiranë: Akademia e Shkencave, 2006.

Çabej 2014 — E. Çabej. Studime etimologjike në fushë të shqipes [Etymological studies in Albanian]. Bleu 5: K–M. Përg. per botim B. Çabej, A. Omari, S. Mansaku. Tiranë: Botime Çabej, 2014.

Çabej 1987/2017 — E. Çabej. Studime etimologjike në fushë të shqipes [Etymological studies in Albanian]. Bleu 3: C–D. Përgat. S. Mansaku dhe A. Omari. Ribot. anastatik. Tiranë: Akademia e Shkencave, 2017. (First published as: E. Çabej. Studime etimologjike në fushë të shqipes [Etymological studies in Albanian]. Bleu 3: C–D. Tiranë: Akademia e Shkencave e RPS të Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, 1987).

de Leeuw, Mair 2009 — J. de Leeuw, P. Mair. Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software*. 2009. Vol. 31. Iss. 3. P. 1–30. Available at: http://www.jstatsoft.org/v31/i03/ (accessed on 16.12.2019).

Demiraj 2013 — Sh. Demiraj. Gjuha shqipe dhe historia e saj [Albanian and its history]. Botim i dytë i ripunuar. Përg. për botim B. Demiraj. Tiranë: Onufri, 2013.

Desnitskaya 1968a — A. V. Desnitskaya. Albanskiy yazik i ego dialekty [Albanian and its dialects]. Leningrad: Nauka, 1968.

Desnitskaya 1968b — A. V. Desnitskaya. Slavyanskie zaimstvovaniya v albanskom yazyke [Slavic loanwords in Albanian]. Moscow: Publishing House of the Academy of Sciences of the USSR, 1968.

Desnitskaya 1987 — A. V. Desnitskaya. O stilisticheskoy funktsii turtsizmov v albanskoy poezii [On the stylistic function of Turkish loanwords in the Albanian poetry]. A. V. Desnitskaya. *Albanskiy yazik i albanskaya literatura* [Albanian language and Albanian literature]. Ed. by V. P. Neroznak. Leningrad: Nauka, 1987. P. 269–276.

Dizdari 2005 — T. N. Dizdari. Fjalor i orientalizmave në gjuhën shqipe [Dictionary of Orientalisms in Albanian]. Rreth 45.000 fjalë me prejardhje nga gjuhët turke, arabe dhe perse. Tiranë: Instituti Shqiptar i Mendimit dhe i Qytetërimit Islam, 2005.

Domosiletskaya 2002 — M. V. Domosiletskaya. Albansko-vostochnoromanskiy sopostavitelnyy ponyatijnyy slovar. Skotovodcheskaya leksika [The Albanian-Eastern

Romance comparative conceptual dictionary. Stock raising vocabulary]. St. Petersburg: Nauka, 2002.

Dunnington 2017a — D. Dunnington. rosm: Plot Raster Map Tiles from Open Street Map and Other Sources. R package version 0.2.2. 2017. Available at: https://CRAN.R-project.org/package=rosm (accessed on 16.12.2019).

Dunnington 2017b — D. Dunnington. prettymapr: Scale Bar, North Arrow, and Pretty Margins in R. R package version 0.2.2. 2017. Available at: https://CRAN.R-project.org/package=prettymapr (accessed on 16.12.2019).

Dyen et al. 1992 — I. Dyen, J. B. Kruskal, P. Black. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society*. 1992. Vol. 82. Part 5. P. iii–132.

Gjinari 1989 — J. Gjinari. Dialektet e gjuhës shqipe [Dialects of Albanian]. Tiranë: Akademia e Shkencave e RPS të Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, 1989.

Gjinari et al. 2007 — J. Gjinari, B. Bahri, Gj. Shkurtaj, Xh. Gosturani. Atlasi dialektologjik i gjuhës shqipe [Dialectological Atlas of Albanian Language]. Vol. 1. Tiranë: Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë; Napoli: Università degli Studi di Napoli L'Orientale, Dipartamento di Studi dell'Europa Orientale, 2007.

Gjinari et al. 2008 — J. Gjinari, B. Bahri, Gj. Shkurtaj, Xh. Gosturani. Atlasi dialektologjik i gjuhës shqipe [Dialectological Atlas of Albanian Language]. Vol. 2. Tiranë: Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë; Napoli: Università degli Studi di Napoli L'Orientale, Dipartamento di Studi dell'Europa Orientale, 2008.

Gjinari, Shkurtaj 2000 — J. Gjinari, Gj. Shkurtaj. Dialektologjia [Dialectology]. Tiranë: ShBLU, 2000.

Goebl 1993 — H. Goebl. Dialectometry: A Short Overview of the Principles and Practice of Quantitative Classification of Linguistic Atlas Data. R. Köhler, B. B. Rieger (eds.). *Contributions to Quantitative Linguistics*. Dordrecht: Springer, 1993. P. 277–315.

Haarmann 1972 — H. Haarmann. Der lateinische Lehnwortschatz im Albanischen. Hamburg: Buske, 1972.

Helbig 1903 — R. Helbig. Die italienischen Elemente im Albanesischen. (Jahresberichte des Instituts für Rumänische Sprache X). Leipzig: Johann Ambrosius Barth, 1903.

Jokl 1934 — N. Jokl. Slaven und Albaner. *Slavia*. 1934. Ročnik XIII. S. 281–325.

Landi 1989 — A. Landi. Gli elementi latini nella lingua Albanese. Napoli: Edizioni Scientifiche Italiane, 1989.

Lindstedt 2000 — J. Lindstedt. Linguistic Balkanization: contact-induced change by mutual reinforcement. D. G. Gilbers, J. Nerbonne, J. Schaeken (eds.). *Languages in Contact*. Amsterdam, Atlanta, GA: Rodopi, 2000. P. 231–246.

Matras 2005 — Y. Matras. The classification of Romani dialects: A geographic-historical perspective. B. Schrammel, D. W. Halwachs (eds.). *General and applied Romani linguistics*. Munich: Lincom Europa. 2005. P. 7–26.

Mair et al. 2015 — P. Mair, J. de Leeuw, P. J. F. Groenen. Multidimensional Scaling in R: SMACOF. Available at: https://klevas.mif.vu.lt/~tomukas/Knygos/MDS2.pdf (accessed 25 April 2020).

Mihăescu 1966 — H. Mihăescu. Les éléments latins de la langue albanaise. *Revue des études sud-est européennes*. 1966. Tome IV. No. 1–2. P. 5–33. No. 3–4. P. 323–353.

Morozova, Rusakov 2018 — M. S. Morozova, A. Y. Rusakov. Chernogorsko-albanskoe yazykovoe pogranichie: v poiskakh "sbalansirovannogo yazykovogo kontakta" [Montenegrin-Albanian linguistic border: In search of "balanced language contact"]. *Slověne*. 2018. Vol. 7. No. 2. P. 258–302. DOI: 10.31168/2305-6754.2018.7.2.10.

Nerbonne et al. 1999 — J. Nerbonne, W. Heeringa, P. Kleiweg. Edit distance and dialect proximity. D. Sankoff, J. Kruskal (eds.). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Stanford: CSLI Press, 1999. P. v–xv.

Nerbonne, Wieling 2018 — J. Nerbonne, M. Wieling. Statistics for Aggregate Variationist Analyses. Ch. Boberg, J. Nerbonne, D. Watt (eds.). *The Handbook of Dialectology*. Hoboken, NJ: John Wiley & Sons, 2018. P. 400–414.

Orel 1998 — V. Orel. Albanian etymological dictionary. Leiden; Boston; Köln: Brill, 1998.

Ölberg 1972 — H. Ölberg. Untersuchungen zum indogermanischen Wortschatz des Albanischen und zur diachronen Phonologie aufgrund des Vokalsystems. Hrsg. von B. Demiraj. (Albanische Forschungen 35). Wiesbaden: Harrassowitz Verlag, 1972.

Papahagi 1963 — T. Papahagi. Dicționarul dialectului aromân, general și etimologic. Dictionnaire aroumain (macédo-roumain), général et étymologique. București: Ed. Academiei RPR, 1963.

R Core Team 2019 — R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2019. Available at: https://www.R-project.org/ (accessed on 16.12.2019).

Rusakov 2013 — A. Y. Rusakov. Nekotorye izoglossy na albanskoy dialektnoy karte (k voprosu o vozniknovenii i rasprostranenii balkanizmov albanskogo yazyka) [Some isoglosses on the Albanian dialect map (towards the issue of development and expansion of Balkanisms in Albanian)]. V. V. Ivanov (ed.). *Issledovaniya po tipologii slavianskikh, baltiiskikh i balkanskikh yazykov (preimuschestvenno v svete yazykovykh kontaktov)* [Studies on typology of Slavic, Baltic, and Balkan languages (mainly in the light of language contacts)]. St. Petersburg: Aleteia, 2013. P. 113–174.

Rusakov 2017 — A. Y. Rusakov. Albanian. M. Kapović (ed). *The Indo-European Languages*. Second edition. New York: Routledge, 2017. P. 552–608.

Rusakov, Morozova 2017 — A. Y. Rusakov, M. S. Morozova. Linguistic complexity: What do Albanian dialects show? Paper presented at the Workshop on Balkan linguistics "Macro and micro variables across the Balkans". 26 October 2017, University of Helsinki, Finland.

Rusakov, Morozova 2018 — A. Y. Rusakov, M. S. Morozova. Linguistic complexity and (micro-)areal history: The case of Albanian. Paper presented at the 51st Annual Meeting of the Societas Linguistica Europaea. 29 August — 1st September 2018, Tallinn University, Estonia.

Rusakov et al. 2018 — A. Y. Rusakov, M. S. Morozova, M. A. Ovsjannikova. Linguistic complexity and lexicon of Albanian dialects: An attempt of quantitative analysis.

Paper presented at the Workshop “First step towards an interactive map of Balkan linguistic features”. 26–27 November 2018, University of Zurich, Switzerland.

Russo 2016 — M. Russo. Differences and interactions between scientific and folk biological taxonomy. P. Juvonen, M. Koptjevskaja-Tamm (eds.). *The Lexical Typology of Semantic Shifts*. De Gruyter Mouton, 2016. P. 493–531.

Séguy 1971 — J. Séguy. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*. 1971. Vol. 35. P. 335–357.

Selishchev 1931 — A. M. Selishchev. Slavyanskoe naselenie v Albanii [Slavic population in Albania]. Sofia: Izdanie Makedonskogo nauchnogo instituta, 1931.

Sobolev 2012 — A. N. Sobolev. Slavische Lehnwörter in albanischen Dialekten. B. Demiraj (ed.). *Aktuelle Fragestellungen und Zukunftsperspektiven der Albanologie. Akten der 4. Deutsch-Albanischen kulturwissenschaftlichen Tagung “50 Jahre Albanologie an der Ludwig-Maximilians-Universität München” (23.–25. Juni 2011, Gut Schönwag bei Wessobrunn).* Wiesbaden: Harrassowitz, 2012. S. 215–232.

Svane 1992 — G. Svane. Slavische Lehnwörter im Albanischen. Aarhus: Aarhus University Press, 1992.

Topalli 2017 — K. Topalli. Fjalor Etimologjik i Gjuhës Shqipe [Etymological dictionary of Albanian]. Durrës: Jozef, 2017.

Vătăşescu 1997 — C. Vătăşescu. Vocabularul de origine latină în limba albaneză din comparaţie cu româna [The vocabulary of Latin origin in Albanian in comparison with Rumanian]. Bucureşti: Vavila Edinf., 1997.

Ylli 1997 — Xh. Ylli. Das slavische Lehngut im Albanischen. 1. Teil. Lehnwörter. (Slavistische Beitrage 350). München: Verlag Otto Sagner, 1997.