Diatopic Patterning of Croatian Varieties in the Adriatic Region

Abstract

The calculation of aggregate linguistic distances can compensate for some of the drawbacks inherent to the isogloss bundling method used in traditional dialectology to identify dialect areas. Synchronic aggregate analysis can also point out differences with respect to a diachronically based classification of dialects. In this study the Levenshtein algorithm is applied for the first time to obtain an aggregate analysis of the linguistic distances among 88 diatopic varieties of Croatian spoken along the Eastern Adriatic coast and in the Italian province of Molise. We also measured lexical differences among these varieties, which are traditionally grouped into Čakavian, Štokavian, and transitional Čakavian-Štokavian varieties. The lexical and pronunciational distances are subsequently projected onto multidimensional cartographic representations. Both kinds of analyses confirmed that linguistic discontinuity is characteristic of the whole region, and that discontinuities are more pronounced in the northern Adriatic area than in the south. We also show that the geographic lines are in many cases the most decisive factor contributing to linguistic cohesion, and that the internal heterogeneity within Čakavian is often greater than the differences between Čakavian and Štokavian varieties. This holds both for pronunciation and lexicon.

1. Introduction

One of the most popular methods applied in traditional geolinguistics (dialectology) is the method of isoglosses, in which areas characterized by different realizations of a single feature are separated by a line – an isogloss. Bundles of such lines were traditionally considered the most important criterion for the division of geolinguistic space into linguistic areas. Despite the tendency to rely on the application of this method in traditional dialectology, even there it has long been recognized that isoglosses do not determine dialectal areas unambiguously because they rarely coincide completely. The isogloss method needs additional assumptions to account for transitional zones and/or dialect continua, even though these are widely recognized to be as common as tightly-knit and readily definable linguistic areas (Chambers & Trudgill, 1998:97).

Brozović, who is aware of the problem, argues that in the case of Croatian, because of specific features of the dialectological make-up of this language, the use of traditional isogloss method is nevertheless sometimes justified: "In our linguistic territory we often find the kind of clear-cut dialectal boundaries that older dialectologists could only dream of; these boundaries occur with intense, clear and dense bundles of isoglosses, whereas it has long been clear to dialectologists that such 'ideal' dialectal boundaries are not a common occurrence in language." (1970:9)[1]. It is our opinion, however, that the division of the Croatian language area into dialect groups is still problematic. This is because although clear-cut dialectal boundaries might be found often in Croatia, they are by no means the rule as Brozović (1970) suggests later on in the paper and as our analysis further down will show. An additional problem which is relevant to Croatian is that migrations have led to geographic splits in formerly uniform areas, which makes the selection of features for isoglosses and the resulting partitioning of varieties into dialect areas absolutely crucial, and it is naturally not always clear which isoglosses are older or more important in genetic terms. These problems often result in a lack of agreement among dialectologist regarding the boundaries of single dialects and even groups of dialects and their coverage, which was an additional incentive to analyze them in the way that would not prioritize only certain structural features in the process of dialectal mapping.

[1] Translated by S.M.Dickey.

It was not until the 1970s that Séguy, the main author of *Atlas linguistique de la Gascogne*, laid the foundations of dialectometry and succeeded in overcoming some of the problems inherent to the method of isoglosses. He based his classificatory work in dialectology on counting the differences, i.e. presence vs. absence of single – phonemic, morphological, syntactic and/or lexical – features between two adjacent varieties in a larger set of dialectological material. The number of differences between two varieties was then expressed as a percentage and was used as an index for calculating the linguistic distance between two locations (Séguy, 1973). The linguistic variation found in this way could subsequently be projected onto a geographic map. After Séguy's pioneering work, Goebl, the editor-in-chief of *Atlas Linguistique de l'Italie et de la Suisse Méridionale*, broadened the application of dialectometry, refined it by adding weighting measures, and developed various techniques for cartographic projection of linguistic diversity (for an overview, see Goebl, 2006). Kessler (1995) was the first to apply the Levenshtein distance algorithm to calculate the distances between the varieties of Irish, a technique which has proven to be a reliable tool for measuring linguistic distances based on pronunciation (see Nerbonne and Heeringa, 2010, for a recent overview). The application of computational methods in general and the Levenshtein algorithm in particular to large amounts of diverse dialectological data has shown that these methods, when applied systematically to extensive data collections, can supplement and improve existing analyses of diatopic variation by systematizing methods, attending to all available data and removing one source of arbitrariness in analysis, viz. the selection of features for isoglosses (see Heeringa & Nerbonne, 1999; Bolognesi and Heeringa, 2002; Gooskens and Heeringa, 2004; Nerbonne and Siedle, 2005; Prokić et al., 2009; Valls et al., in press; Nerbonne, 2009, etc.). The new techniques provide a new way of accounting for the relation between diatopic variation and geography (Nerbonne et al., 2008; Nerbonne, 2011), language contact effects (Nerbonne et al., 2010), and contribute to the study of diffusion (Nerbonne, 2010) and intelligibility (Gooskens et al, 2010).

1.1. A short review of earlier scholarship on diatopic variation in Croatia

The Croatian linguistic area[2] consists of approximately 23 dialects (Brozović, 1970:6-7) that can roughly be split into three main dialect groups[3] – Štokavian, Čakavian and Kajkavian – the names of which are derived from the form of the interrogative and relative pronoun meaning 'what' (*što*, *ča* and *kaj* respectively) and which putatively correspond to relatively well-definable, even if linguistically not quite homogeneous areas prior to the 15th century. However, an extremely diversified geographic terrain consisting of numerous natural boundaries, important political and cultural borders that have crisscrossed these regions throughout history, and – most of all – large-scale migrations from the southeast in the wake of the expansion of the Ottoman Empire between the 14th and 16th century, caused significant changes in the diatopic landscape of the region. The last of these factors is considered mainly responsible for the decrease in the number of dialects, for the loss of contact zones which previously marked the transitions between the three groups of dialects, as well as for the formation of numerous linguistic enclaves (Brozović, 1970). Moreover, the heavy 'overlaying' of various adstrates (for a graphic representation, see Brozović, 1970:18-19) caused by migrations, vexes the question of how to identify the features characteristic of the three dialect groups (Vermeer, 1982:279-289; Lisac, 2009:17). This complexity and the absence of proof that Kajkavian, Štokavian, and Čakavian derive from three distinct proto-varieties (especially in the case of the latter two) is the reason why some dialectologists even doubt the usefulness of this three-fold classification of Croatian dialects (Vermeer, 1982:279-289; Kalsbeek, 1998:2-5). The truth is that sometimes there is as much diversity within each of these traditionally assumed groups as between them. This holds particularly with regard to the division of the Čakavian and Štokavian groups since they form neither clearly

[2] The question whether or not Croatian forms part of a larger dialect continuum and if so, how we should call that continuum, is not relevant within the scope of this article, since the language varieties analyzed here are all spoken within the Croatian Republic and the speakers all refer to themselves as Croats and to their language as Croatian.

[3] In Croatian dialectology the groups are normally referred to as *narječja*; for terminological differences between *narječje*, *grupa dijalekata*, *dijalekt* and other terms, see Brozović 1960. The names used to refer to any of these (groups of) varieties will be capitalized throughout the paper, but the terms used to refer to one specific feature in one or more of those groups of varieties (e.g. cakavism, (i)(j)ekavism) or even the varieties preserving such a feature (e.g. cakavian, (i)(j)ekavian) will not be capitalized, following the model which uses lower case to refer to "r-less varieties".

delineated dialectal areas nor a continuum in the real sense of the word (Brozović, 1970). The terms Štokavian, Kajkavian and Čakavian are nonetheless well-established in Croatian dialectology and provide the basis for a lot of dialectological work – even by those who are not very supportive of their division (e.g. Vermeer). For this reason and to simplify the discussion we will continue to refer to these three broadly defined groups of dialects. In this way we hope to facilitate the comparison of our results with those of earlier dialectological studies.

In this paper the focus will be on the analysis of some of the varieties found along the Adriatic coast. For the most part this region originally belonged to the Čakavian dialectal area. Today, however, it is far from being homogeneous and is mostly characterized by intermingling and overlaying of various Čakavian and Štokavian adstrates mostly due to migrations and the more recent influence of neo-Štokavian Standard Croatian.[4]

One of the simplest divisions of the Čakavian dialect group distinguishes the more peripheral and conservative north-western Čakavian on the one hand, and the innovative south-eastern Čakavian group, on the other. The latter group can then be further divided into more conservative coastal varieties and more innovative insular varieties of SE Čakavian (Ivić, 1981). Vermeer (1982) proposes a classification of Čakavian dialects into three groups based on the presence of the neocircumflex (secondary lengthening of short stressed vowels which resulted in long falling intonation, e.g. *gȋne* 'perish, 3sg.present', *stâri* 'old'), and the reflex of the Proto-Slavic front vowel *ě (jat): a) NW Čakavian characterized by the neocircumflex and different reflexes of jat (Kalsbeek, 1998:7); b) Central Čakavian without neocircumflex, but with a specific ekavian-ikavian reflex of the PS *ě according to Meyer-Jakubinskij's Law,[5] and with many further innovations shared with Kajkavian and SE Čakavian dialects; and c) SE Čakavian with an ikavian (jekavian on the Island of Lastovo) reflex of the PS *ě, and with other

[4] The two Štokavian dialects that were in close contact with Čakavian and are thus relevant in the context of this paper are the Western neo-Štokavian ikavian dialect (see Lisac, 2003:50-76) and the East-Herzegovinian neo-Štokavian (i)jekavian dialect (ibid.98-120). The term neo-Štokavian is used for a group of Štokavian dialects that differ from the old Štokavian ones in terms of accentuation. Neo-Štokavian dialects were affected by a backward shift of all accents in words that were originally accented on any but the first syllable. Neo-štokavian serves as a dialectal basis of the Croatian standard variety.

[5] According to Meyer-Jakubinskij's Law PS */ě/ > /ɛ/ occurred in restricted environments only, namely where *ě was followed by a back vowel {a, o, u, y, ъ} after { t, d, n, l, r, s, z }. In all other cases */ě/ > /i/ (Jakubinskij, 1925).

innovative features shared with many Štokavian dialects. This classification has been adapted and further elaborated by Kalsbeek (1998).

The classification of the Čakavian area primarily on the basis of the reflex of PS *ě, and secondarily on the basis of consonantal criteria and accentuation, has been adopted by Lisac (Lisac, 2009:30-31) who classifies Čakavian dialects into: 1) the Buzet dialect in which PS *ě has been partly preserved in the form of a closed ę /e/; 2) the Southwestern Istrian dialect in which PS *ě > /i/; 3) North Čakavian marked by PS *ě > /ɛ/; 4) Middle Čakavian in which PS *ě > /i/ or /ɛ/ (see note 3); 5) South Čakavian with PS *ě > /i/; and 6) the Lastovo dialect in which PS *ě > /(i)jɛ/. The two ikavian dialects, Southwestern Istrian and Southern Čakavian, are distinguished on the basis of a consonantal criterion, namely the conservation of the consonantal group -šć- /ʃtɕ/ found in South Čakavian (in words such as *šćap* 'stick'*, dvorišće* 'yard') while -šć- /ʃtɕ/ > -št- /ʃt/ in the Southwestern Istrian dialect (*štap, dvorište*) (Lisac, 2009:30).

1.2. Earlier dialectometric work in Croatia

Dialectometry has not been widely used in Croatian dialectological scholarship. However, a version of a computationally based calculation of linguistic distances was applied to analyze the degree of differentiation mostly between highly concentrated local vernaculars in small geographically delimited areas such as single islands or peninsulas (Sujoldžić et al., 1982/83; Sujoldžić et al., 1986; Sujoldžić et al., 1988; Sujoldžić, 1989; Sujoldžić et al., 1989; Sujoldžić, 1990; Sujoldžić et al., 1990; Sujoldžić et al., 1992/93) and sometimes to compare the results obtained in several such areas (Sujoldžić, 1994; Sujoldžić, 1997; Szirovicza et al., 1997). These studies were performed within the context of holistic anthropological research on the Eastern Adriatic where differences and similarities in local speech variants were used as an indicator of the socio-cultural micro-differentiation of particular subpopulation groups. In most cases the analysis compared data by assigning different categories to lexical units that differed from others on the basis of prosodic, phonological or lexical features. The linguistic distances were calculated using Hamming measure of similarity that determines the percentage of

congruity between lexical units (Sujoldžić et al. 1986). Factor analysis was applied to some of the lexical variables (Sujoldžić, 1994), and Hidden Markov Models were trained to classify local varieties into the 'Čakavian' or 'Štokavian' type at a rate of 85% correctness (Szirovicza et al., 1997). While the major purpose of the studies above was not to discuss language as such but to use it to study population change by migration, they provided some useful synchronic evidence of the indigenous dialect patterns of the Eastern Adriatic area. Generally, the area of the 'pure' old Čakavian dialect in the investigated region has become significantly reduced under the influence of migrational movements. Thus, the comparative study of the Middle Dalmatian islands has shown that in this region it is not possible to define a sharp Čakavian-Štokavian boundary, although the Štokavian influence drops noticeably from east to west (Sujoldžić, 1997).[6]

2. Aim of the study

The aim of this study is to calculate linguistic distances among different speech varieties spoken in the Adriatic region, to analyze the (large) set of distances for natural groups, and then to compare the groups we detect to some of the dialectological classifications and lexicostatistical analyses carried out before. We were interested in finding out whether and to what extent synchronous similarities and differences among the investigated varieties would correspond to dialectological classifications based primarily on diachronic criteria. In order to do that, it was also important to find out whether the effects of external factors, such as language contact, would be discernible in this kind of dialectometric analysis. Supposing that the effects of language contact are manifested differently at different linguistic levels, we were also interested in comparing the differences between linguistic distances found on the basis of pronunciation and lexical

[6] The intermixing of greater or lesser numbers of autochthonous and immigrant populations resulted in Štokavian with a stronger or weaker Čakavian substrate, while towards the west Čakavian is spoken with a stronger or weaker Štokavian superstrate. Dialectal differentiation was also influenced by the sharp separation between town and country that brought about significant differences in the application of foreign (Venetian) elements between country and town. While the towns were always to some degree bilingual (so that foreign elements suffered fewer changes), the rural areas were strictly monolingual (Croatian) and accepted foreign elements only indirectly through the towns, trying to adapt them as much as possible to their own speech system. The manifestations of that adaptation were quite varied, reflecting the differences in the specific local speeches (subdialects).

analysis. We used as a leading hypothesis the postulate that the lexicon changes more easily than pronunciation (Thomason and Kaufmann, 1988), even though the quantitative evidence of this has been found lacking (Spruit et al., 2009).

Although there is an overlap in the database used for this study and the lexicostatistic and earlier dialectometric analyses (Sujoldžić, 1994, 1997; Šimičić, 2005), the present study is more comprehensive in the number of locations included, even if the number of analyzed lexical items is not as high as in some earlier studies. This study is also innovative with respect to Croatian dialectology as it uses Levenshtein distance to calculate the Croatian pronunciation distances. For the most part, the analysis was carried out by means of Gabmap (http://www.gabmap.nl), a web-based version of the L04 program designed and implemented by the Groningen dialectometry group (Nerbonne et al., 2011).

3. Sample and research methodology

Before discussing the results, a short description of the sample, a few remarks on the digitization of the database, and an overview of the statistical methods used in the analysis is given.

3.1. Sample description

The linguistic data were collected in 88 settlements located in what is traditionally considered the Čakavian dialectal area of Croatia (for the geographic position of Croatia, see Fig. 1a). The area investigated extends along the east Adriatic coast from the Istrian peninsula in the north-west to Pelješac, the second-largest Croatian peninsula in the south-east and comprises eight islands (Krk, Pag, Silba, Olib, Brač, Hvar, Korčula, Vis), two peninsulas (Istra and Pelješac) as well as three coastal settlements along the Makarska Coast (Fig.1b and 1c). Three Croatian-speaking villages in the Italian province of Molise were also included in the analysis as the variety used there largely originated in the coastal area investigated. However, due to the centuries of isolation and intense contact with Italian dialect varieties it has become something of an isolate among the Croatian dialects.

Figure 1a. The shaded area is Croatia. The present study focuses on the coastal varieties found in Fig. 1b and 1c.

The data were collected during field trips undertaken in the period from 1978 to 2003 by A. Sujoldžić for the majority of locations, except for the island of Vis where she was joined by another researcher. Additionally, data from Istria were taken from the questionnaire for the Croatian Linguistic Atlas.[7] Two Čakavian dialectologists, P. Šimunović and B. Finka, specializing in the varieties under investigation were involved in the phonetic transcription of the lexical data and its preparation for further dialectological and statistical analysis, which ensured both quality and a high degree of consistency in the transcription of different varieties. Most of the data used in the analysis has been published.[8]

[7] Originally *Upitnik za srpsko-hrvatski dijalektološki atlas*, the use of which had been kindly permitted in 1978 by the then Language Department of the Institute of Philology and Folklore. Today, the questionnaires form part of the database for the Croatian Linguistic Atlas at the Institute for the Croatian Language and Linguistics.

[8] The data collected on the island of Pag in Sujoldžić et al., 1990; for Krk in Sujoldžić et al., 1992/93; for Silba and Olib in Sujoldžić, 1989; for Brač in Sujoldžić et al., 1988; for Hvar in Sujoldžić et al., 1982/83; Korčula in Sujoldžić et al., 1986; Pelješac in Sujoldžić et al., 1989; Molise in Sujoldžić 1990. The exceptions are the data for the Makarska Coast and the most recently collected data from the Island of Vis, which are archived at the Institute for Anthropological Research.

Figure 1b. The map of the northern Adriatic area. Only the sites included in the analysis are indicated.

Figure 1c. The map of the locations investigated in the southern Adriatic area.

The main idea behind the collection of the data used here was to probe genuine speech instead of infrequently used and obsolescent forms. This meant that fieldworkers noted different lexical items elicited without insisting specifically on archaic dialectological features. This implies that not all the criteria characteristic of traditional dialectology were observed. But we emphasize that the data were collected by the same team in a consistent manner, meaning that the responses are comparable and that the data indicated where variation exists. A questionnaire was used as the primary means of eliciting lexical items, the use of which was then sometimes additionally checked in spontaneous speech.[9]

Several respondents were interviewed wherever available. The respondents were generally adults (aged 20 to 80) originating from the settlement under investigation, who had lived elsewhere for not more than six months and were exposed to standard Croatian in the course of their formal schooling for as little as possible but never more than 12 years.[10] The age and sex of the respondents were not crucial factors in choosing interviewees. All different realizations for every single concept occurring in each settlement were recorded in writing and taken into consideration in the analysis.

The selection of words was originally based on the so-called basic vocabulary word list (Swadesh, 1952) subsequently adapted to the specifics of the Čakavian dialect as well as to cultural and historical factors in the populations under investigation. As such it has already proven to be a suitable linguistic (lexical) dataset to trace linguistic microevolution of settlements in small and relatively isolated but linguistically diversified areas along the Adriatic coast (Sujoldžić et al., 1982/83; Sujoldžić et al., 1986; Sujoldžić et al., 1988; Sujoldžić, 1990; Sujoldžić et al., 1992/93). Although the original lists usually contained over 100 basic vocabulary items, for the present research we use only 92. We selected these 92 words because they were registered in all 88 localities, with the exception of *glad* 'hunger' and *zjenica* 'eye pupil, which were not recorded in Molise

[9] In Croatian dialectological research the use of questionnaires has another advantage, namely that of eliciting nouns in the nominative singular, which would be extremely difficult if spontaneous speech were the only means of data collection and a relatively high number of lexemes from different speakers was needed.

[10] Obviously it is increasingly difficult nowadays to find respondents who have never moved from their village of origin and/or who have not completed at least a few years of secondary school. This is also why on the Island of Vis, where the last fieldwork was conducted, the average level of education of our respondents was generally higher compared to the speakers we interviewed twenty or more years earlier.

and *mast* 'fat', which had to be disregarded in all Istrian localities due to a morphologically different form recorded there (instrumental instead of nominative). In addition, special attention was paid that all variants of a single concept appear in the same form, normally nominative singular for nouns and infinitive for verbs.

Due to different rates of change in phonological and lexical inventories, the former being generally subject to stricter constraints than the latter, we analyzed the two levels separately. The entire 92-item word list (Table 1) was used in the lexical analysis since etymologically diverse lexemes might indicate the influence of linguistic contact, for example, through the spread of cultural influences. However, for the analysis at the phonological level it was necessary to eliminate from the original list those concepts for which significant lexical variation had been found (items in bold in Table 1). In the phonological analysis only those 84 items were considered for which a cognate variant appeared in over 85% of the locations, which was sufficient to carry out a reliable analysis (Cronbach $\alpha = 0.95$, local incoherence = 1.87). All the lexemes in the table are given first in the Standard Croatian orthography with marked accentuation, followed by the IPA transcription, and the most basic word meaning in English. Please note that English glosses for all the words are provided in Table 1 below; for this reason they are omitted in the text.

A number of distinctive dialectal features were attested in the 84 concepts analyzed at the level of pronunciation. The presence, absence or specific combinations of these features have been taken to indicate important dialectal divisions in traditional classifications, which presumably contribute to synchronic linguistic distances, although not necessarily to the same extent. Some of these phonetic characteristics discussed in earlier literature include:[11]

a) Accentuation

The types of accents and their distribution have been taken into consideration as every transcribed word was marked for accent (tone or, alternatively, stress when tone was

[11] The examples are meant to give a rough overview of the possibilities of the realization of a specific feature mentioned and not of all the possible realizations of a single lexeme. This is also the reason why no accentuation is provided for the examples. To avoid misunderstanding, we emphasize that stress differences were part of the measurements.

irrelevant, as well as length where applicable).[12] The strings were aligned so that each stressed vowel and each vowel marked by lengthening was regarded as a separate segment, i.e. units consisting of a phonetic symbol and diacritical mark(s) were the bases of comparison with other units in sequence (word) comparison (see Table 2). In many cases accentuation can be regarded as the most indicative feature distinguishing Čakavian and neo-Štokavian linguistic strata.

b) Reflex of the Proto-Slavic *ě, which can be ekavian, ekavian-ikavian, ikavian, (i)jekavian (e.g. PS *sěme > *seme, sime, sjeme*; PS *tělo > *telo, tilo, tijelo*); [13]

c) Syllabic r /r̩/ (e.g. PS *pьrstъ > *parst, prst*)

d) Reflexes of PS *ę as a /a/ or e /ɛ/ after j /j/, č /ʧ/, ž /ʒ/ (e.g. PS * (j)ęzykъ > *jazik, jezik*);

e) Reflex of PS *tj as a palatal affricate ć /tɕ/, palatodental affricate ć̑ /ʧʲ/, or as a palatal plosive ť /tʲ/ (e.g. PS *kǫtja > *kuć̑a, kuća, kuťa*; PS *čьlověkъ > *čovjek, čo̍vjek*);

f) Closing and/or diphthongization of long vowels, if any (e.g. PS *golva > *glava, glạva, glaōva, glova*; PS *nosъ > *nos, nọs, nuōs, nus*);

g) Cakavism, i.e. the reduction of two phonemic sets č /ʧ/, š /ʃ/, ž /ʒ/ and c /ʦ/, s /s/, z /z/ to one c /ʦ/, ṡ /ɕ/, ż /ʑ/ (e.g. *čovik > covik; san > ṡan; žena > żena*);

h) Weakening and/or reduction of consonants (e.g. lenition as in PS *dъno < *dno, lno*, and complete reduction as in PS *dъkt'i > *kći* > ći, *ćer;* PS > *pčela* > *čela;* PS *sъglobъ > *zglob* > *zlob, zlub;* PS *dolnь > *dlan* > *lan, lon*);

i) Retention of final l /l/ (e.g. PS *pepelъ > *pepeo, pepel*);

j) Apocope in the infinitive form (e.g. PS *sъpati > *spati, spat, spa*);

k) Delateralization ĺ /ʎ/ > j /j/ (e.g. PS *žuljь > *žuĺ* > *žuj*) and depalatalization ĺ /ʎ/ > l /l/ (e.g. *žuĺ* > *žul*);

---

[12] Vowels can be short and long in Croatian. It is also characterized by pitch accent, which means that stressed vowels carry either falling or rising tone resulting in four possible types of accents: short falling (marked by double grave ̏), long falling (marked by inverted breve ̂), short rising (marked by grave accent ̀), and long rising (marked by acute accent ́). In addition, if unstressed vowels are long, the macron ̄) is used to indicated non-tonic length.

[13] *Hrvatski jezični portal* (http://hjp.srce.hr/), see footnote 12, was also the source for the Proto-Slavic etymologies cited.

l) Presence or absence of h /x/ in the phonemic inventory (e.g. PS *kruxъ > *kruh, kruv*);

m) Prothesis of j /j/ (e.g. PS *usta > *usta, justa*);

n) Dissimilation of liquids /r – r/ > /l – r/ (e.g. PS *rebro > *rebro, lebro*);

o) Metathesis (e.g. *lakat* / *latak*);

p) Devoicing of final voiced consonants (e.g. *zub* > *zup*; *nož* > *noš*);

q) Apophony (e.g. PS *grobъ > *grob*, *greb*).

TABLE 1

All the concepts analyzed are rendered in Standard Croatian. Non-bold items were analyzed on the basis of pronunciation only (section 4.1), while all 92 items (incl. those in in bold) were analyzed lexically (section 4.2.). [14]

1	***bèdro* /ˈbědro/ ‘thigh’**	32	*làkat* /ˈlâkat/ ‘elbow’	63	*rèbro* /ˈrěbro/ ‘rib’
2	*cvijȅt* /tsviˈjê:t/ ‘flower’	33	*lâž* /ˈlâ:ʒ/ ‘lie’	64	*rúka* /ˈrǔ:ka/ ‘arm’
3	*čèlo* /ˈʧělo/ ‘forehead’	34	*líce* /ˈlĭ:tse/ ‘face’	65	*sȁn* /ˈsân/ ‘dream’
4	*čòvjek* /ˈʧôvjek/ ‘man’	35	*lȉšće* /lˈî:ʃtɕe/ ‘leaves’	66	*sèstra* /ˈsěstra/ ‘sister’
5	*díjete* /ˈdĭ:jete/ ‘child’	36	*màgla* /ˈmăgla/ ‘fog’	67	*sȋn* /ˈsî:n/ ‘son’
6	*dȉm* /ˈdîm/ ‘smoke’	37	*mâjka* /ˈmâ:jka/ ‘mother’	68	*sȉsa* /ˈsîsa/ ‘breast’
7	***djȅd* /ˈdjêd/ ‘grandfather’**	38	*mâst* /ˈmâ:st/ ‘fat’	69	*sjèkira* /ˈsjěkira/ ‘ax’
8	***djèvōjka* /ˈdjěvo:jka/ ‘girl’**	39	**mèđāš /ˈměʤa:ʃ/ ‘boundary stone’**	70	*sjȅme* /ˈsjême/ ‘seed’
9	*dlȁn* /ˈdlân/ ‘palm’	40	mjèhūr /ˈmjêhu:r/ ‘bubble’	71	*spávati* /ˈspă:vati/ ‘to sleep’
10	*dnȍ* /ˈdnô/ ‘bottom’	41	mjèra /ˈmjêra/ ‘measure’	72	*sȗnce* /ˈsû:ntse/ ‘sun’
11	*glâd* /ˈglâ:d/ ‘hunger’	42	*mjȅsec* /ˈmjêsets/ ‘moon’	73	*sȕša* /ˈsû:ʃa/ ‘drought’
12	*gláva* /ˈglă:va/ ‘head’	43	*mlàdīć* /ˈmlăditɕ/ ‘young man’	74	*sùza* /sˈûza/ ‘tear’
13	*gnȏj* /ˈgnô:j/ ‘pus’	44	*mlijéko* /ˈmliječko/ ‘milk’	75	*tijȇlo* /ˈtijê:lo/ ‘body’
14	*grȍb* /ˈgrôb/ ‘tomb’	45	*mȕž* /ˈmû:ʒ/ ‘husband’	76	*ùho* /ˈûho/ ‘ear’
15	*ȉme* /ˈîme/ ‘name’	46	*nòga* /ˈnǒga/ ‘leg’	77	*ùmrijēti* /ˈŭmrije:ti/ ‘to die’
16	*jáje* /ˈjă:je/ ‘egg’	47	*nȍkat* /ˈnôkat/ ‘nail’	78	*ùnuk* /ˈŭnuk/ ‘grandchild’
17	*jȅsti* /ˈjêsti/ ‘to eat’	48	*nȏs* /ˈnô:s/ ‘nose’	79	*ústa* /ˈŭ:sta/ ‘mouth’
18	*jèzik* /ˈjězik/ ‘tongue’	49	*nȏž* /ˈnô:ʒ/ ‘knife’	80	*ȗš* /ˈû:ʃ/ ‘louse’
19	***kȁmēn* /ˈkâme:n/ ‘stone’**	50	*njȅdra* /ˈɲêdra/ ‘bosom’	81	*vȁtra* /ˈ vâtra/ ‘fire’
20	*kćȉ* /ˈ ktɕî:/ ‘daughter’	51	*ȍbrva* /ˈ ôbrva/ ‘eyebrow’	82	*vráta* /ˈ vră:ta/ ‘door’
21	*kȉša* /ˈ kîʃa/ ‘rain’	52	*ȍko* /ˈ ôko/ ‘eye’	83	*zglȏb* /ˈ zglôb/ ‘joint’
22	*kòbila* /ˈ kǒbila/ ‘mare’	53	*òtac* /ˈ ǒtats/ ‘father’	84	*zjènica* /ˈ zjěnica/ ‘pupil (eye)’
23	***kȏla* /ˈ kôla/ ‘cart’**	54	*óvca* /ˈ ǒ:vtsa/ ‘sheep’	85	*zȗb* /ˈ zû:b/ ‘tooth’
24	*kòljeno* /ˈ kǒʎeno/ ‘knee’	55	***pȁs* /ˈ pâs/ ‘dog’**	86	*zvijézda* /ˈ zvijě:zda/ ‘star’
25	*kȏnj* /ˈ kôɲ/ ‘horse’	56	*pčèla* /ˈ pʧěla/ ‘bee’	87	*žâr* /ˈ ʒâ:r/ ‘ember’
26	*kȏst* /ˈ kô:st/ ‘bone’	57	*pȅpeo* /ˈ pêpeo/ ‘ash’	88	*žèna* /ˈ ʒěna/ ‘woman’
27	*kòšulja* /ˈ kǒʃuʎa/ ‘shirt’	58	*plèća* /ˈ plětɕa/ ‘shoulders’	89	*žȅnsko* /ˈ ʒênsko/ ‘female’
28	*krȏv* /ˈ krôv/ ‘roof’	59	***prȋšt* /ˈ prî:ʃt/ ‘pimple’**	90	*živjeti* /ˈ ʒĭ:vjeti/ ‘to live’
29	*krȕh* /ˈ krûx/ ‘bread’	60	*pȓsa* /□pɻ□sa/ ‘chest’	91	*žȋvanj* /ˈ ʒɻ□:vaɲ / ‘grindstone’
30	*krȋv* /□kɻ□:v/ ‘blood’	61	*pȓst* /ˈ pɻ□st/ ‘finger’	92	*žûljˈ* /ˈ ʒû:ʎ / ‘blister’
31	*kȕća* /ˈ kûtɕa/ ‘house’	62	*pût* /ˈ pû:t/ ‘way’		

[14] The source used for the Standard pronunciation was *Hrvatski jezični portal* (http://hjp.srce.hr) compiled on the basis of various lexicographic publications including: *Rječnik hrvatskoga jezika* by Anić (I.ed. 1991, II ed. 1994, III ed. 1998), *Pravopis hrvatskoga jezika* by Anić and Silić (2001), *Veliki rječnik hrvatskoga jezika* by Anić (2003), *Rječnik stranih riječi* by Anić-Goldstein (I ed. 1998, II ed. 2000), and *Hrvatski enciklopedijski rječnik* by Anić et al. (2003).

It can be seen from the list that only a certain number of features pertaining to the phonological and prosodic levels were analyzed, while morphological features, which had not been collected in large quantity, were largely ignored. The method of data collection, emphasizing words, meant that syntactic features are also absent from the material.

3.2. Digitization of the database

All the data were rendered in traditional Croatian dialectological orthography in the source materials. Because Gabmap operates best either with X-Sampa or Unicode IPA transcriptions, a special program was written to convert all the symbols used in the database into Unicode IPA symbols. Table 2 contains all the tokens that appear in the database and that are treated as separate segments in pronunciation analysis. The conversion tables are based on the Croatian IPA standard (Landau et al., 1999:66-69) where applicable, and otherwise on the solutions made by the authors. Because IPA does not provide the means of rendering all possible pronunciations, in a few cases the closest alternative symbol was used to designate the pronunciation (e.g. /ɕ/ and /ʑ/ to denote subdental fricatives, and not alveolopalatal fricatives, which did not appear in our material). The same IPA standard was considered most suitable to represent four different accents found in standard Croatian (short falling, long falling, short rising, and long rising, see footnote 10). For other types of accents, a different notation was used. Short stressed vowels were marked by means of the primary stress symbol, while the primary stress symbol followed by length was used to mark the Čakavian acute (Table 2). Vowels making up diphthongs were treated as separate segments because in earlier studies treating them as single tokens did not alter the aggregate results significantly (Heeringa, 2004:174).

TABLE 2
Conversion table into IPA[15]

Consonants

Cr	IPA	Cr	IPA	Cr	IPA	Cr	IPA	Cr	IPA	Cr	IPA
b	b	t’	tʲ	g	g	l	l	p	p	t	t
c	ts	d	d	h	x	Ĺ	ʎ	r	r	v	ʋ
č	ʧ	đ	ʤ	j̇	j̇	m	m	s	s	z	z
ć	ʧʲ	d’	dʲ	j̇	j̇	n	n	ṡ	ɕ	ż	ʑ
ć	tɕ	f	f	k	k	ń	ɲ	š	ʃ	ž	ʒ

Vowels

Cr	IPA	Cr	IPA	Cr	IPA	Cr	IPA	Cr	IPA	Cr	IPA
a	a	e	ɛ	i	i	o	ɔ	u	u	ṛ	ṛ
à	â	è	ê	ì	î	ò	ô	ù	û	r̀	r̂
â	âː	ê	êː	î	îː	ô	ôː	û	ûː	r̂	r̂ː
à	ǎ	è	ě	ì	ǐ	ò	ǒ	ù	ǔ	r̀	ř
á	ǎː	é	ěː	í	ǐː	ó	ǒː	ú	ǔː	ŕ	řː
ã	ˈaː	ẽ	ˈɛː	ĩ	ˈiː	õ	ˈɔː	ũ	ˈuː		
ā	aː	ē	ɛː	ī	iː	ō	ɔː	ū	uː	ə	ə
ˈa	ˈa	ˈe	ˈɛ	ˈi	ˈi	ˈo	ˈɔ	ˈu	ˈu	ə̀	ô
ạ	ɑ	ẹ	e	y	ɪ	ọ	o	ü	ɨ	ô	ôː
ậ	ɑ̂ː	ệ	êː	ỳ	î	ộ	ôː	ű	ɨ̂ː	ˈə	ˈə
ạ́	ǎː	ẹ́	ěː	ŷ	îː	ọ́	ǒː	ű	ɨ̌ː		
ą̄	ɑː	ę̄	eː			ǭ	ˈoː	ű̃	ˈɨː		
ą̃	ˈɑː					ǭ	oː	ü̃	ˈɨː		
ä	æ										
a͡	æ̂										

3.3. Computational and statistical analysis

Levenshtein distance (also known as “edit distance”) was used to calculate the linguistic distances on the basis of pronunciation between the locations listed above. It is an algorithm used to treat sequence (string) data which sums the “editing” costs of rewriting one string into another by means of insertions, deletions and substitutions. Because it is

[15] This table is not meant to be an exhaustive list of segments used in Croatian dialectology, but includes only those that appeared in our material (database). It does not correspond entirely to the standardized version for the conversion of the symbols traditionally used in Croatian dialectology into IPA as proposed by the Institute of Croatian Language and Linguistics as far fewer symbols appear in the database used for the purposes of this analysis than actually appear in the Croatian Linguistic Atlas. A more or less standardized IPA representation has been available for Standard Croatian only since Landau et al. (1999).

usually possible to rewrite one string into another in different ways, often yielding different costs, Levenshtein distance is defined to be the least sum of the costs needed to transform one string into another. This means that, even though the mapping of /deˈvîtʃina/ into /divojˈʧʲina/ (or vice versa) for instance, can be performed as shown in the left column of Figure 2, only the 'translation' in the right column corresponds to Levenshtein distance.

devîtʃina	(insert i)	1			
dievîtʃina	(delete e)	1			
divîtʃina	(insert o)	1	devîtʃina	(substitute i for e)	1
divoîtʃina	(delete î)	1	divîtʃina	(insert o)	1
divotʃina	(insert j)	1	divoîtʃina	(substitute $ʧ^{j}$ for ʧ)	1
divojtʃina	(substitute î for i)	1	divoîtʃʲina	(substitute j for î)	1
divojtʃina	(substitute ʧ for $ʧ^{j}$)	1	divojtʃʲina	(substitute î for i)	1
divojtʃʲina			divojtʃʲina		
7 / 9 = 0.78			**5 / 9 = 0.56**		

Figure 2. There are different ways of mapping the string /deˈvîtʃina/ into /divojˈʧʲina/. While the mapping in the left column is possible, only the one on the right can be considered Levenshtein distance as it is the least costly mapping. In both cases the result has been normalized by dividing the cost of transformation by the number of tokens in the string in order to discount the effect of word length to the overall cost of transformation. For simplicity, the diacritic sign for long falling accent was not assigned a separate value in this and in the following figure.

Another way of understanding the Levenshtein distance is to examine the matrix which the algorithm uses to keep track of costs incrementally. Note that the substitution of segments (i,j) costs 1 if the aligned segments are different, and 0 if they are the same. Insertions and deletions are both assigned the cost '1'. Then the recurrence for row i and column j is:

$$d(i,j)= \min (d(i-1,j-1)+subst(i,j), d(i-1,j)+1, d(i,j-1)+1)$$

This corresponds to postulating a substitution, an insertion or a deletion, respectively, at the i,j cell. This way of calculating the distance guarantees that what is computed is the minimal distance needed to map one string onto another (Fig.3).

		d	**i**	**v**	**o**	**j**	**ʧʲ**	**ɲ**	**n**	**a**
	0	1	2	3	4	5	6	7	8	9
d	1	**0**	1							
e	2	1	**1**	2						
v	3		2	**1**	2					
ɲ	4			2	**2**	3	4			
ʧʲ	5				3	**3**	**4**	5		
i	6					4	4	**5**	6	
n	7							5	**5**	6
a	8								6	**5**

Figure 3. Matrix used in calculating the distance between the two realizations of the word *djevojka* 'girl'. The number that appears in the last cell is the minimal distance needed to map the one string to the other, while the numbers in bold indicate the minimal distance between the corresponding alignments.

The linguistic motivation for applying Levenshtein distance rather than a simpler string edit distance measure lies in the fact that in allowing insertions and deletions as well as substitutions, it is more likely to account for the process of linguistic differentiation as it is perceived.[16] Hamming distance, for instance, does not calculate the least cost of transformation, but compares segments one by one often yielding an unrealistically high distance between two phonemic realizations (Fig. 4) so that the resulting distances measured by Levenshtein and Hamming can differ to a great extent (HD = 100 % and LD = 40%) (Fig. 4).[17]

Hamming distance:	Levenshtein distance:
ʑ l ɔ̂ b	ʑ l ɔ̂ b
z g l ɔ̂ b	z g l ɔ̂ b
1 2 3 4 **5 / 5 = 1**	1 **2** **/ 5 = 0.4**

Figure 4. Comparison of the distances calculated by Hamming and Levenshtein measure respectively. In both cases the cost of transformation was divided by the number of segments in the longer string.

[16] Levenshtein distance is not always an ideal measure for calculating linguistic distances. Because the version we use operates with one symbol at a time it cannot recognize and treat adequately segment changes which involve changes of place of different segments such as different types of metathesis (e.g. contact metathesis as in *žlica / lžica* 'spoon' or distant metathesis as in *gomila / mogila* 'crowd' or *lakat / latak* 'elbow'). In all cases of metathesis the cost of transformation is 2 provided everything else remains the same, which is an unrealistically elevated cost of transformation compared with the actual historical process.

[17] As one referee noted, Hamming distance is often defined to be applicable only to (measure the number of differences between) strings of equal length. But there is a natural generalization, which is used in Fig.4. which counts the number of differences in the strings when they ar e aligned on the left plus the difference in the length of the two. Kruskal (1999: 1) and Gusfield (1997:403) define it this way. These are also excellent technical introductions to sequence comparison.

A linguistic constraint is applied in processing to ensure that vowels may only align with vowels, consonants only with consonants, while semi-vowels and syllabic consonants can be aligned with either vowels or consonants. The simplest version of the Levenshtein algorithm was used in which aligning identical phones yields a cost of 0 and the difference between non-identical ones always costs 1 regardless of their phonetic similarity. There are more sophisticated versions in which 'editing' costs depend on phonetic similarity (Heeringa, 2004). Although it is clear that the difference between /p,k/ is smaller than that between /p,g/, which are in turn more similar than /p,z/, it has been shown that even elaborate feature-based segment distances, besides being somewhat arbitrary from the point of view of their importance in the perception of speakers/hearers, do not contribute significantly to *overall* average distances between varieties (Heeringa, 2004:186, 194). The quantity of the data compensates for the roughness of the measure.

In order to account for specific accentual features relevant in the analysis of the Croatian dialects, the standard VC alignment was slightly modified by using a specially designed 'user-defined' string edit distance. The definition treats the stress as simply a modifier (diacritic) on a vowel. Every insertion and deletion of any modifier is assigned the cost of 0.2.[18] In our analysis accentual features are cumulative in that the distance between two different vowels, say /ɑ/ and /ɔ/ with different accentuation is larger than the distance between two different vowels with same accentuation or two identical vowels accented differently (Fig. 5):

[18] Croatian accent consists of three dimensions – stress (ictus), tone and length. Tone (if relevant) can appear only on stressed segments. Initially, all stressed vowels were marked for both stress and tone using separate symbols. In some other types of processing based on edit distance stress is counted as a separate segment, so that a different placement of stress requires two operations, one to insert a stress in the new position and one to delete it from the old. Needing two operations to model single changes is inelegant and results in unrealistically high edit costs for this single change. The role of the tone, represented by a diacritic was often underestimated in such an analysis on the other hand. Since tone can be just as indicative of the overall character of a certain variety as stress is, the two are represented in comparable ways – both as diacritics, and the user-defined feature definition assigns similar costs to changing, deleting or inserting the diacritics. Length, too, was treated in this way. Despite the fact that the value of 0.2 assigned to each prosodic feature may seem arbitrarily low, the fact that all differences in prosody add up (substitution function does not apply to prosody) means that the total difference assigned to prosody in two strings can amount to 1 or, in extremely rare cases even somewhat higher (consider the difference between 'kratkosílazn̄i' vs. 'kratkoš̄ilāzni', and other examples, in Lupić, 2001:88).

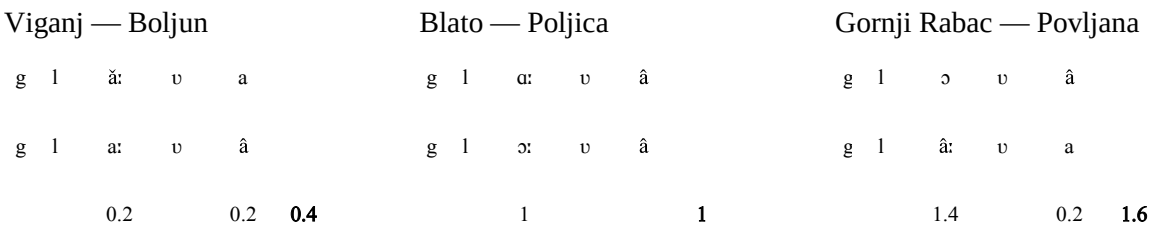

Figure 5. The alignments of different variants of the same word are shown. In the first example the influence of different tone on the overall distance is shown and in the second the additional effect of a different vowel.

Lexical distances were calculated on the basis of 'cognateness' by designating all etymologically cognate lexemes of each concept with the same number. The numbers were subsequently compared categorically.[19]

We compare nearly four thousand (more precisely, (88 × 87)/2 = 3828) pairs of sites both with respect to the lexicon and with respect to pronunciation. In both comparisons the differences per item are summed and the mean distance of all items present was calculated as the difference between the sites. It is best to imagine two site × site matrices in which the (i,j) cell represent the lexical or pronunciational distance between site i and j. Naturally, we do not need to fill in more than one half of the matrix, as distance(i,j) = distance(j,i).

The most commonly used methods for the analysis of linguistic distance matrices are cluster analysis and multidimensional scaling (MDS). Cluster analysis is a statistical method used to group elements into clusters on the basis of their similarity – however similarities or differences are defined. MDS is used to represent the relation of elements in a low-dimensional space on the basis of the distances calculated between them. Because both clustering and MDS are analyses of the differences in the dataset, they simplify and never completely reflect original measurements in their full complexity. However, previous studies have noted that three-dimensional MDS representation usually accounts for about 90% of the variation in the distance matrix and can thus be considered reliable (Heeringa, 2004; Prokić & Nerbonne, 2008). If too little variation is represented

[19] Binary comparison could have been used instead if there had not been multiple answers.

in an MDS analysis, this will be obvious in a low correlation between distances in the input matrix and distances in the inferred two- or three-dimensional solution.

All clustering techniques are more problematic than multidimensional scaling as they may assign single items (varieties) to different clusters based on very small differences, which means that even insignificant alterations in the distance matrix could change the groupings altogether. Although Ward's method, Group average and Weighted group average are generally more reliable than other clustering techniques, their instability becomes clear when the clusters obtained in this way are projected to MDS plots. Regardless of the chosen technique, the varieties are sometimes assigned to given clusters arbitrarily. For this reason traditional clustering will be omitted altogether in the paper and probabilistic or 'noisy' clustering (Nerbonne et al., 2008) will be used instead. It is based on adding different levels of noise to the data and repeatedly calculating in order to assess the stability of individual clusters. The levels of added noise can vary, but we have opted for a 0.2 threshold, which corresponds to 20% of a standard deviation in the data. Based on probabilistic clustering, composite (noisy) cluster maps are obtained – maps based on the superimposition of many maps obtained during many iterations of clustering using different random amounts of noise each iteration.

4. Results

The results will be presented in two parts. First, the results of the analysis based on pronunciation (phonological and prosodic) differences will be given, followed by the discussion on lexically based distances. The findings of the two analyses will be compared in the last section of the paper.

4.1. Analysis of pronunciation differences

The representation of pronunciational differences based on the analysis of average Levenshtein distances for 84 cognate words indicates that the dialects in the north-western Adriatic region are less homogeneous than the ones from the south-eastern

Adriatic region (Fig. 6a).[20] This can be inferred from the fact that in the northwest only a few neighboring varieties are connected by dark lines which indicate phonetic and prosodic similarity:[21] the eastern Krk varieties containing the speech of Omišalj, Dobrinj, and Vrbnik; southern Istrian varieties of Rakalj, Medulin, and to a lesser extent Rovinjsko Selo; south-eastern Pag varieties of Vlašići, Dinjiška and Povljana and its northwestern varieties of Lun, Novalja and Kolan; and the varieties spoken on the islands of Silba and Olib. However, the whole northern part of Istria and to a somewhat lesser extent central and western Istria appear to be extremely diverse in that the adjacent varieties resemble each other only slightly. Figure 6b shows that these varieties remain linguistically dissimilar to other Istrian and in particular to other NW Adriatic varieties. For instance, the Buzet dialect is often considered a separate group, but the differences within that group are in many cases greater than those between the varieties that supposedly belong to a completely different groups of dialects, namely Čakavian and Štokavian ikavian, in SE Adriatic region (Fig. 6b).

There is significantly less internal diversity among the varieties in the southeastern Adriatic region (Fig. 6b). This region is thus characterized by larger groups of similar varieties. One of these groups includes almost all the varieties on the island of Brač excluding only two on the southern part of the island (Milna and Bol) while several varieties in the east form a separate group. Other such groups include the whole western half of the island of Hvar (exluding the cakavian town of Hvar on the west); the Pelješac varieties of Kuna, Potomje and Pijavičino; and a group of varieties on the western part of the island of Korčula (Vela Luka, Blato, Smokvica and Čara).

The local varieties of Croatian spoken in the Italian region of Molise show conspicuous similarity among themselves, but the centuries of isolation have contributed to the increase of the linguistic distance between these varieties and those they were closely related to prior to the migration overseas.

[20] Most of the locations used for this study are densely distributed on very small patches of land, mostly islands. In order to make visual representation of data possible a so-called 'disperse' function was created in Gabmap. This means that the locations are marked by a dot, while a pointer indicates further information, such as coloring assigned to a particular variety. This method of representation helps to distinguish geographically close locations visually.

[21] The results of the regression analysis of linguistic and geographical distances in the northern and southern Adriatic region, $r^2 = 0.08$ and $r^2 = 0.36$ respectively, indicate that more variance is explained by geography for the data in the southern Adriatic.

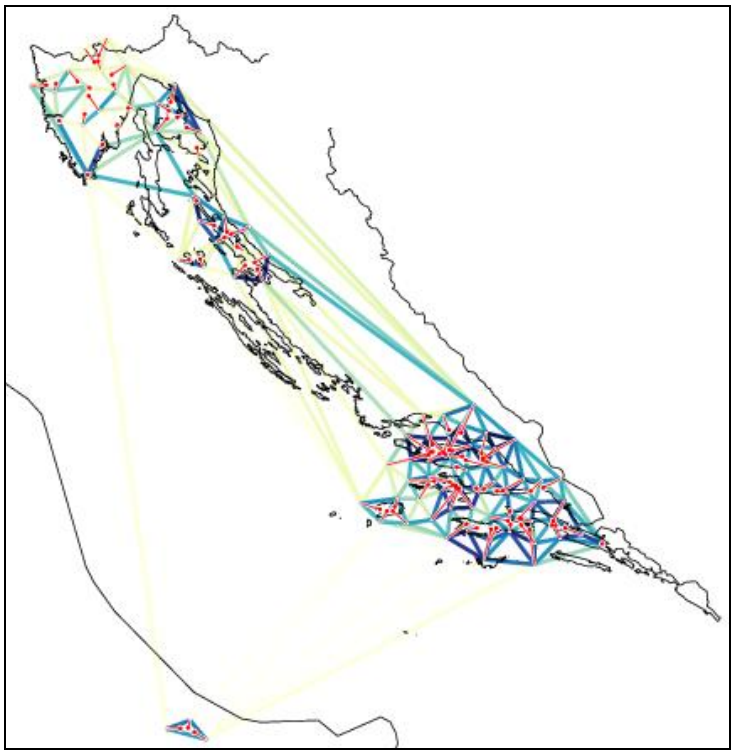

Figure 6a. Network map of the average pronunciation distances based on the dataset containing cognate variants for 84 concepts in 88 locations (Cronbach's α = 0.95). Only adjacent sites are connected by lines. The darker the lines are, the more similarity there is between the varieties. The red lines are auxiliary, pointing to the geographical location of the sample more exactly.

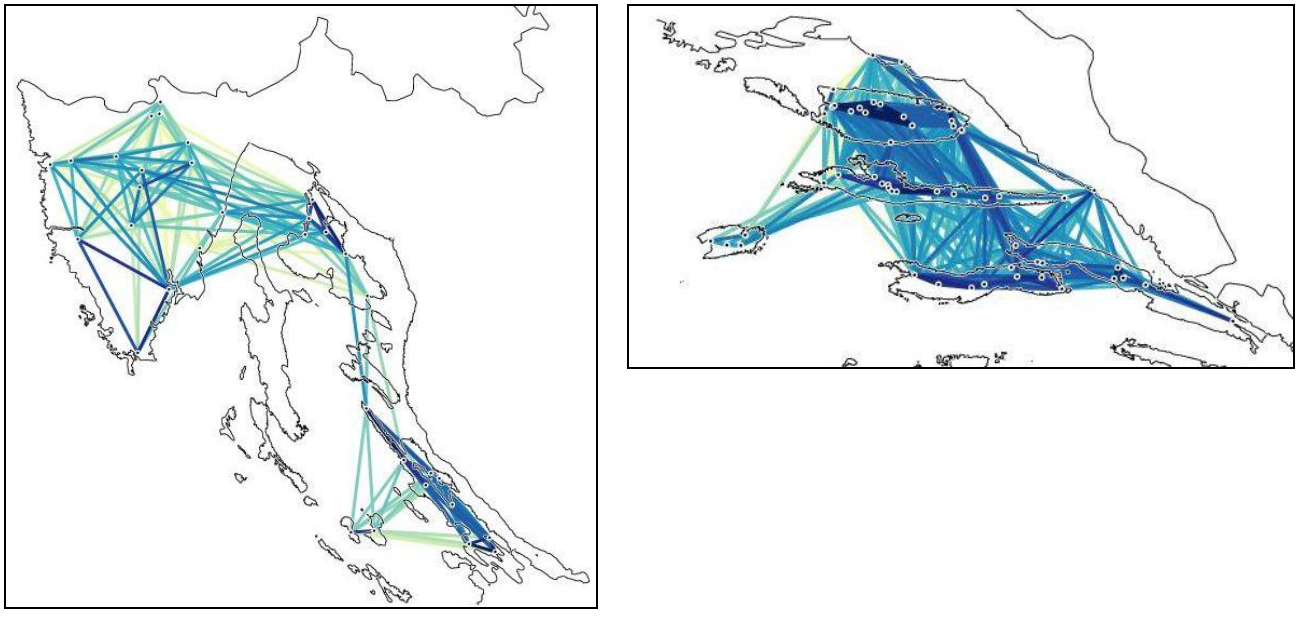

Figure 6b. 'Beam map' of the pronunciation differences between varieties analyzed in the northern (Cronbach α = 0.93) and southern (Cronbach α = 0.95) Adriatic region. Lines are suppressed over larger geographic distances.

4.1.1. Cluster analysis

The results of probabilistic clustering are represented first in a dendrogram (Fig. 7a) and are then projected on the map (Fig. 7b). The dendrogram, based on probabilistic clustering (using 20% noise) using a combination of group average and weighted group average, shows that there are no large stable clusters in the Adriatic region (Fig. 7a). However, a number of smaller-sized clusters were detected with over 60% certainty.

There are five cakavian varieties (see 3.1 sub g)), two in the NW Adriatic (Baška and Pag) and three in the SE Adriatic (Sutivan and Milna on the island of Brač, and Hvar on the island of Hvar). These varieties form one of the most stable clusters (>99% certainty) in the data, in spite of their geographic dispersion. Although cakavism has also been attested in some other varieties – Gornji Rabac in the Labin area in Istria and on the island of Vis – other phonological specifics such as a different reflex of PS *ě, different accentual patterns, or less consistent application of cakavism caused them to remain detached from the main cakavian cluster.

While the speech of Gornji Rabac, which is also characterized by cakavian pronunciation to some extent, did not cluster with any other varieties, it is noteworthy that the three cakavian varieties on the island of Vis form a separate but very stable cluster (91% certainty). This can at least partly be attributed to the fact that the island of Vis is the most peripheral of all middle Dalmatian islands, which is the reason why quite a lot of specific features can be found there. Also, earlier studies of language change in both real and apparent time have shown that certain dialectological features of the varieties on the island were undergoing change (Šimičić & Sujoldžić, 2009), and one of these features regards the cakavian pronunciation. Although still retained to some extent among older speakers, cakavism has become reduced and partly lexicalized among younger speakers, who tend to preserve the pronunciation of the alveolar affricate c /ts/ instead of its postalveolar counterpart č /ʧ/, while subdental fricatives ṡ /ɕ/ and ż /ʑ/ are virtually absent from their speech. Another feature contributing to the distance of Vis varieties from other cakavian varieties concerns ‘mixed’ and often double accentuation present in some other South Čakavian varieties especially among younger speakers. The archaic Čakavian stress and tone are increasingly influenced by other accentual systems

(primarily neo-Štokavian), which is visible both in the appearance of rising tone, shifted stress and occasionally double accents.[22] The sources of this influence are Standard Croatian and (especially) the urban koine of Split.

Probabilistic clustering has confirmed the heterogeneity of dialectal areas in the north Adriatic region, particularly in Istria. Besides the above mentioned site of Gornji Rabac, which stands out as an isolate in our dataset, the speech of Vabriga in the north-west of Istria also turns out to behave quite differently. Although Vabriga is situated in the region settled by south Čakavian (ikavian) speakers (Lisac, 2009:158), sporadic cakavian-like features (e.g. č /ʧ/ > c /ts/, š,s /ʃ,s/ > ṡ /ɕ/, and ž,z /ʒ,z/ > ż /ʑ/) set it apart from other varieties in the region, including the nearby varieties of Kaldir and Kaštelir. Other south Čakavian varieties in Istria, namely in Medulin, Rakalj, and Rovinjsko Selo form a separate cluster, which is marked by more significant Štokavian influence than other varieties in Istria (Sujoldžić et al., 1990) and is much closer to southern Pag varieties (Povljana, Vlašići and Dinjiška). This SW part of Istria was settled mainly in the 16th century by migrants who originated from the inland part of the Makarska Coast, the region where Štokavian was spoken, but who were exposed to Čakavian linguistic influence in Šibenik and Zadar regions on their way to Istria (Lisac, 2009:62). The island of Pag is originally Čakavian, but has been exposed to the immigration of Štokavian speakers more than some other islands (Sujoldžić et al, 1990:7) throughout its history because of its proximity to the coast. The Štokavian superstrate is more visible in the southern part of the island while Čakavian has been better preserved in the north. The Štokavian influence is visible primarily in neo-Štokavian accentuation (*gláva* 'head', *jèzik* 'tongue'), features such as *dj > đ /ʤ/ (*mèđa* /mêʤa/ 'boundary') and the retention of the final -i in the infinitive as in *živiti* /ʒiːviti/ 'to live'. Although many such features are not exclusively Štokavian they distinguish older Čakavian from a more recent Štokavian adstrate on the island of Pag.

North Čakavian dialects are spoken in the Buzet region and in the central part of Istria. Although both dialects are regarded as native to the Istrian peninsula, they differ considerably with respect to the reflex of PS *ě and *ǫ and thus form two separate

[22] By 'double accentuation' we imply the appearance of two accents of more or less equal strength in a polysyllabic word, usually the old (Čakavian) and the new (neo-Štokavian) accent (Kapović, 2004:101).

clusters. In the Buzet dialect to which the varieties of Brest, Nugla and Sv. Martin belong (72% certainty) we find short stressed PS *ě > ẹ /e/ (*jẹ̀s* 'to eat' in Nugla) and long stressed PS *ě > ẹ /e/ (*tẹ̑lo* 'body'), but also unstressed PS *ě > i /i/ (*mlīkȍ* 'milk'), even though there are many exceptions to these developments (e.g. *človȉk* 'man' in Brest and Nugla or *mȉra* 'measure' in Sv.Martin and Brest).[23] The central Istrian North Čakavian dialect is not always uniform either. Within this dialect the speech of Boljun, Pazin and Žminj forms a very stable cluster (100% certainty) characterized by the reflex of short PS *ě > e /ɛ/ (*mèra* 'measure', *sekȉra* 'ax') and of long PS *ě > ie /ie/ (*tiẽlo* 'body', *mliēkȍ* 'milk') and PS *ǫ > o /ɔ/ (*r*^{u}*ōkȁ* 'arm'), while Brseč and Lupoglav form a separate cluster (76% certainty) due to a different development of Proto-Slavic jat and the nasal vowel: PS *ě > e /ɛ/ (*sème* 'seed', *sekȉra* 'ax'*; mlēkȍ* 'milk', *tẹ̑lo* 'body'), while PS *ǫ > o /ɔ/ or u /u/ (*rokȁ / rukȃ* 'arm').

Except for the cakavian varieties of Baška (Krk) and Pag (Pag), the northern islands of Krk and Pag are much more homogenous linguistically as compared to the diversity found in Istria. Although all the varieties investigated on the island of Krk are characterized by the presence of neocircumflex accent and ekavian-ikavian reflex of PS *ě, only the most archaic Čakavian varieties on the island, those of Omišalj, Vrbnik and Dobrinj, form a separate cluster (100% certainty).There *ъ > e /ɛ/ or o /ɔ/ as in the penultimate syllable of the words *olkъtь > *lȅket / lȍkot* (elsewhere *lȁkat* 'elbow') and *sъnъ > *sȇn / sôn* (elsewhere *sȃn* 'dream').[24] The speech of Dubašnica and Njivice on the western part of the island is more similar to the Pag varieties (89% certainty), many of

[23] In fact, Lisac (2009) distinguishes this dialect from other North Čakavian dialects solely on the basis of the reflex of jat, unlike Vermeer (1982) and Kalsbeek (1998) who group it together with the North Čakavian varieties due to the presence of neocircumflex.

[24] For other features that reflect the archaic character of the speech of Omišalj, Dobrinj and Vrbnik, see Sujoldžić et al. (1992/93:4-5).

which are marked by greater or lesser Štokavian superstrate features brought by the migrants from the coast.[25]

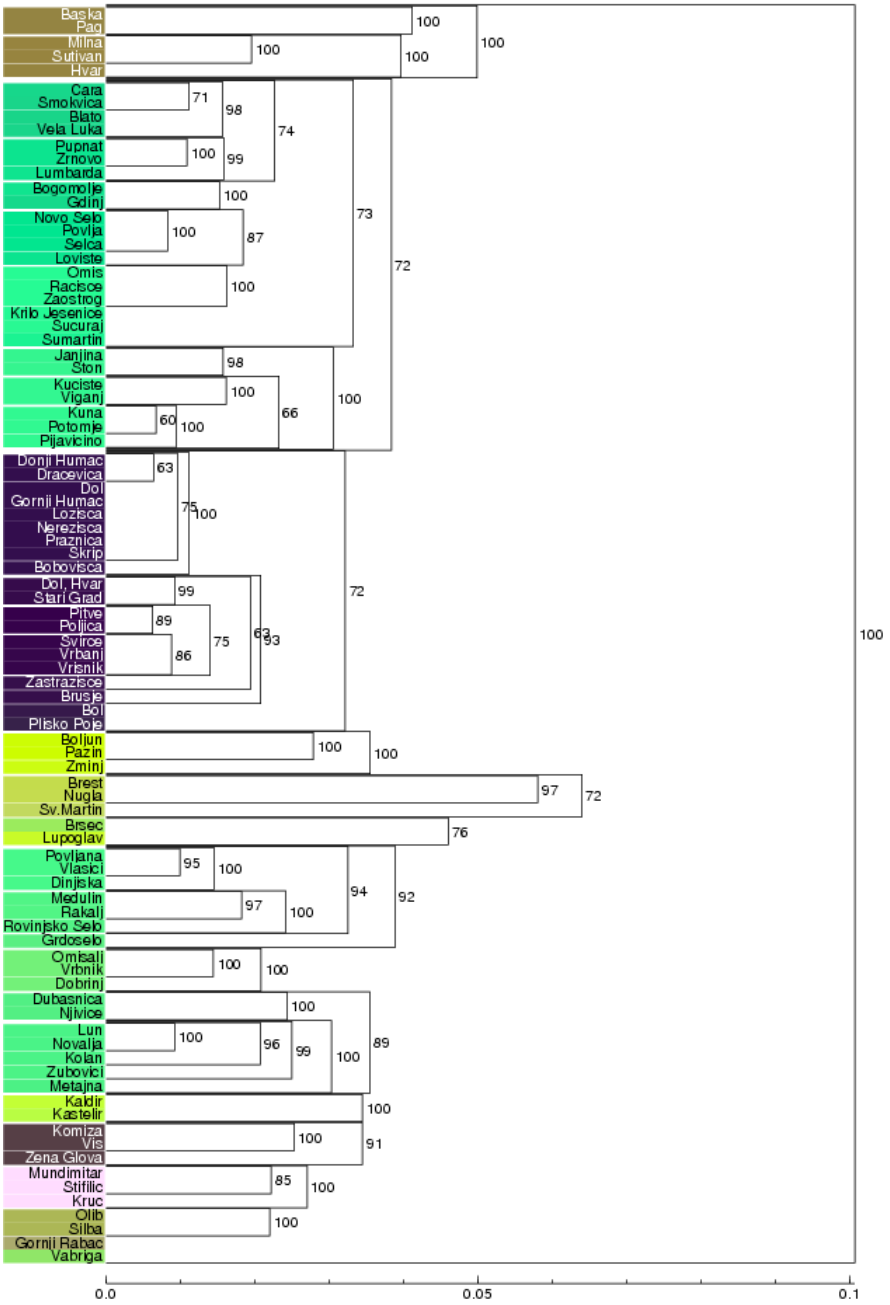

Figure 7a. Probabilistic dendrogram shows the grouping of analyzed varieties on the basis of the combined group and weighted average clustering (noise=0.2sd). The x-axis is the mean normalized Levenshtein distance between varieties, and the numbers to the right of the cluster brackets indicate the probability (percentage of cluster iterations) with which certain varieties will group together.

[25] Again, because of the Central Čakavian ikavian-ekavian reflex of PS *ě, Lisac groups practically the whole island of Krk together with other Central Čakavian varieties (Lisac, 2009), while the presence of neocircumflex in its most archaic varieties qualifies it for the North Čakavian group in Vermeer's (1982) and Kalsbeek's (1998) classification.

Besides Pag, which has been significantly influenced by the language of the migrants from the mainland, other representatives of the Central Čakavian dialect are the varieties spoken on the small and distant islands of Silba and Olib. Due to their distance from the coast, these two islands have been less exposed to external linguistic influences and thus form a cluster separate from other Central Čakavian varieties found on Pag.

In contrast to the linguistic splitting of the dialects in the North, most of the varieties in the south are grouped into two larger clusters. The first one is a relatively coherent cluster made up of quite similar South Čakavian varieties on the islands of Hvar and Brač (72%, purple on map 7b). The variety spoken in Bol, on the southern coast of Brač, is the most different of all Čakavian varieties on the island due to diphthongization, a feature found also in the nearest and northernmost varieties on the neighboring island of Hvar.

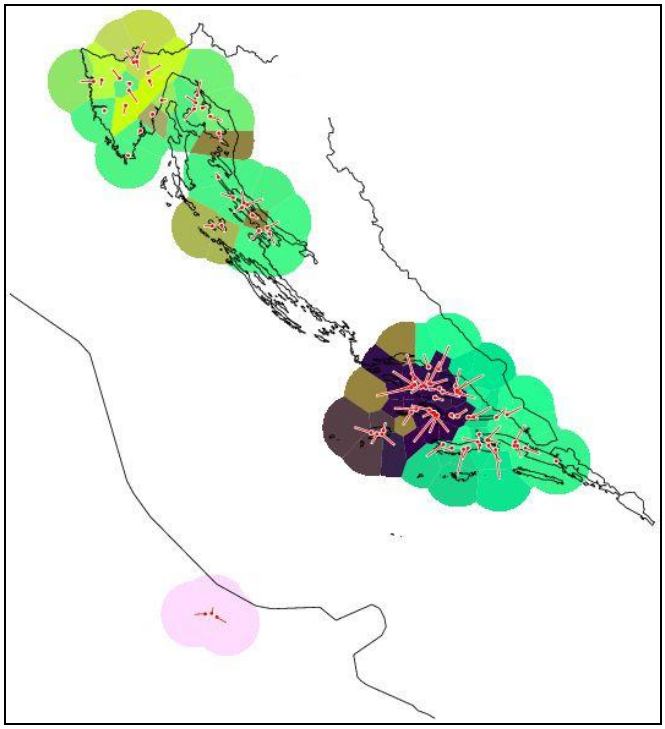

Figure 7b. Noisy cluster map based on group and weighted average probabilistic clustering (noise=0.2sd). This is a cartographic representation of Fig. 7a.

The second cluster found in the data in the southern Adriatic region consists of all the varieties characterized by a greater or lesser degree of Štokavian admixture (72%, dark green). However, the dendrogram based on the analysis of pronunciation differences indicates that the physical (island) boundaries have more effect on the aggregate

pronunciation similarity of the varieties than the relative degrees of Štokavian and Čakavian admixture, which seem to be only of secondary importance (cf. Sujoldžić et al., 1982/83; 1988). An example of this is the situation on Pelješac. All of Pelješac (with the exception of Lovište) forms a separate cluster regardless of the internal differentiation, though sub-clustering is visible on a lower level according to the reflex of PS *ě. The peninsula forms a continuum ranging from jekavian neo-Štokavian varieties of Ston and Janjina in the east, to Štokavian ikavian with some Čakavian influence in the west (*mȅja* vs. *mèđa* 'boundary' on the east), and finally Lovište in which the Čakavian adstrate is felt more than elsewhere (Sujoldžić et al., 1989; Lisac, 2003:98, 2009:139). It is of interest that Lovište on the western coast of Pelješac, which was founded by settlers from the eastern part of the island of Hvar (Bogomolje and Gdinj), is grouped with Čakavian-Štokavian varieties on Brač rather than with those on Hvar (cf. Sujoldžić, 1997: 296). The varieties on the islands of Hvar (Sućuraj), Brač (Sumartin), and Korčula (Račišće), which are normally considered Štokavian, do not form a single unified cluster, although they belong to the same group in a higher-level clustering which includes other 'mixed' Čakavian-Štokavian varieties.

4.1.2. Multidimensional Scaling

Multidimensional scaling (MDS) offers an alternative view of the dialectal data as it provides a more nuanced representation of similarities and dissimilarities of the varieties assigned to separate and well-defined groups by means of clustering. The recognition and visualization of gradual transitions between the groups makes MDS a better means of presenting another dimension of all diatopic variation, namely dialect continua. MDS is also superior to clustering in that it produces more stable results, i.e. results which are much less likely to be influenced by small differences in the input data.

In an MDS visualization (Fig. 8a) it becomes obvious that the Croatian dialect of the Italian province of Molise, although considerably different from any variety spoken on the opposite shore of the Adriatic, most resembles Štokavian-influenced varieties of South Čakavian (Lisac, 2009) spoken on the western part of the Pelješac peninsula as well as the speech of Krilo on the Makarska coast and that of Kolan on the island of Pag.

Regardless of the innovations, such as the reduction of short vowels and the loss of tone on short stressed vowels, these varieties have retained an ikavian reflex of jat as well as characteristic accentuation patterns in those lexemes which were not replaced by Italian loanwords.

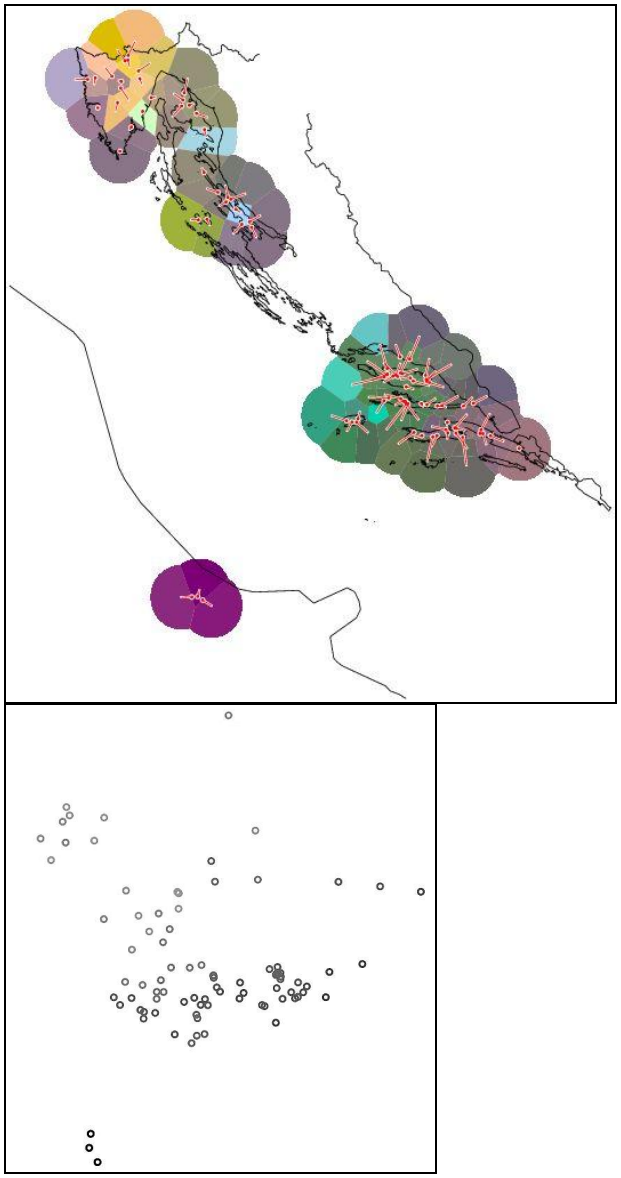

Figure 8a. Multidimensional scaling map in three dimensions mapped onto RGB color space (r=0.91) and a MDS plot (r=0.87). The figure is based on pronunciation differences found in 88 varieties along the Adriatic coast.

A "corrective" role of MDS compared to clustering comes to the fore in the representation of cakavian varieties, especially in view of the town of Hvar as a 'bridge' between the cakavian varieties of Milna and Sutivan on Brač on the one hand and Vis and

Komiža on Vis on the other (Fig.8b). Although the varieties on the island of Vis formed a separate cluster in probabilistic clustering, in Figure 8a it is clear that the Vis varieties form part of the cakavian group in the south. Separate clustering might therefore be a consequence of their peripheral position in the Čakavian continuum. MDS reveals a number of other similarities invisible in clustering, e.g. the similarity of Southwest Istrian (Medulin, Rakalj, Rovinjsko Selo) to the South Čakavian varieties of southern Pag (cf. Lisac, 2009:139), the Štokavian of the Makarska coast (sp. Krilo), and to a lesser extent the varieties on the eastern part of Korčula and western Pelješac (Fig. 8a). All of these varieties are marked by some degree of dialect mixing. It is interesting that they are separated by only very small linguistic distances (ld)[26] despite the fact that they are set apart from each other geographically and that all of them were formed by the overlaying of Čakavian and Štokavian adstrates, but not necessarily in the same order. SW Istrian, for instance, was formed by adding a Čakavian superstrate onto a Štokavian substrate (cf. Lisac, 2009:62), whereas on the islands of Pag, Korčula and Pelješac the linguistic history was reverse (cf. Lisac, 2009:158).

The previously undetected similarity of all the autochthonous Istrian North Čakavian dialects is made much more visible in this kind of dialect data visualization. These dialects are related even though their PS reflexes differ. The differing PS reflexes naturally also contribute to internal diversification visible in Fig 6a and 6b, but they are not substantial enough to counterbalance the aggregate similarity of the cluster.

[26] The distance between Rakalj, Medulin, and Rovinjsko Selo is ld ≤ 0.020, the distance between any of these locations and southern Pag (Povljana, Dinjiška, Vlašići) is ld ≤ 0.030, Omišalj ld = 0.035, and western Pelješac ld ≤ 0.035.

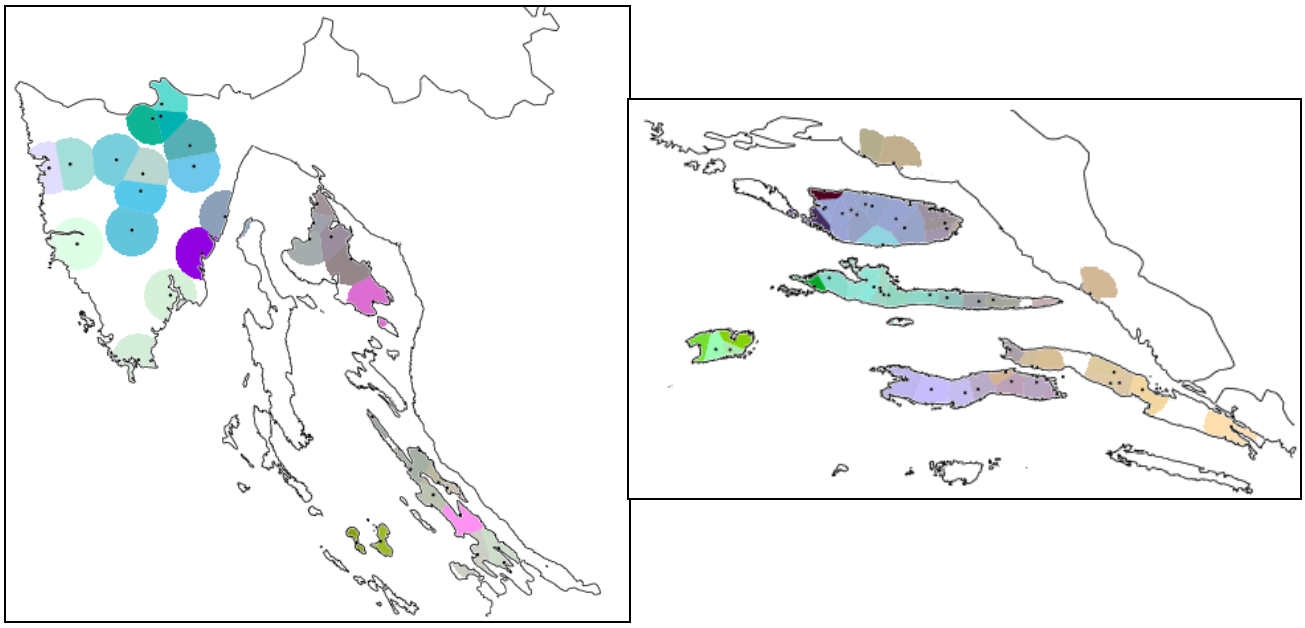

Figure 8b. Classical multidimensional scaling in three dimensions mapped onto RGB color space for the north Adriatic region (r=0.92) and south Adriatic (r=0.96) region.

On the other hand, MDS maps do not obfuscate significant linguistic differentiation. The difference between two different types of Štokavian, the neo-Štokavian ikavian of the Makarska coast and the neo-Štokavian ijekavian of the easternmost part of the peninsula of Pelješac, for instance, is also more visible in the MDS map (8a) than in the map based on clustering (Fig. 7b). The speech of Silba and Olib stands out again as distinct from all other varieties in the MDS representation (see also Fig. 8b). The same is true of the cakavian varieties of Baška (on Krk) and Pag (on Pag). We conjecture that the purple tone in the Gornji Rabac area indicates the presence of cakavism (present very slightly in Vabriga as well), while the presence of blue indicates its similarity to other neighboring Central Istrian North Čakavian varieties.

Figure 8b indicates that the speech of Boljun is very similar to the Buzet dialect and that it is difficult to draw a line between them, while Lupoglav is conspicuously similar to other central dialects, although this could not be inferred from the cluster analysis.

In a similar vein, the centuries-long administrative and cultural ties between the speech of the towns Hvar and Vis are (Škreblin et al., 2002:336) reflected in their linguistic similarity. In a similar vein, linguistic proximity of Bol on the southern coast of Brač and the Čakavian varieties on the island of Hvar (Fig.8b) can be attributed to the

isolation of Bol from other Čakavian locations on the island by Vidova gora, the mountain which forms a natural (physical) boundary to the north, while the sea channel in this case presumably promoted the contact with the nearest part of the island of Hvar on the south. The grayish areas on the easternmost parts of Brač and Hvar indicate a different (neo-Štokavian) influence, although the varieties spoken there really form a continuum with the neighboring Čakavian varieties. The same cannot be said of the cakavian speech of Milna, Sutivan, and Hvar, and their Čakavian neighbors, however.

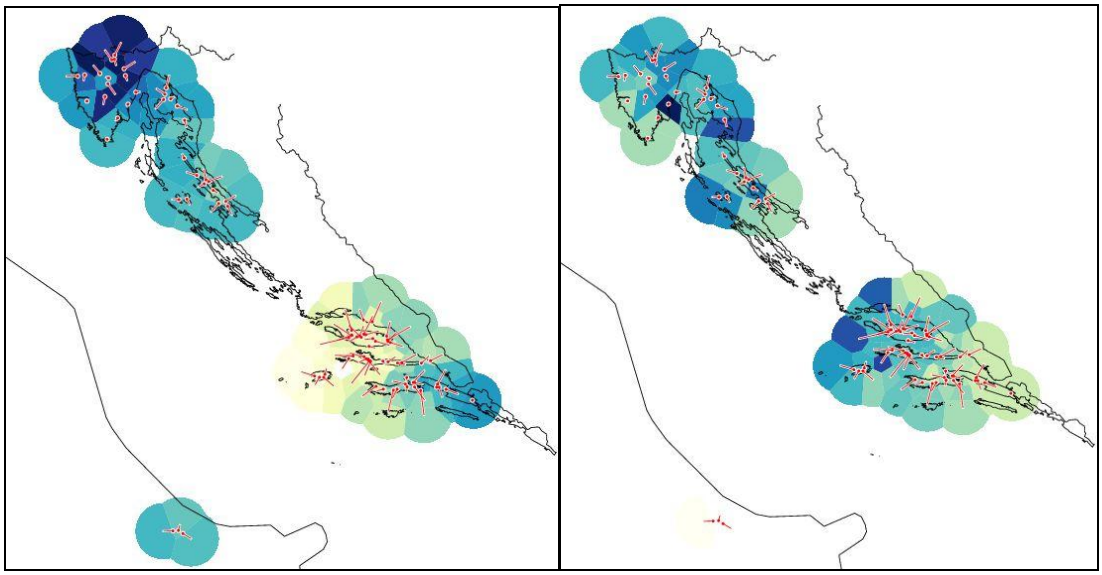

Figure 8c. One-dimensional pronunciation-based MDS maps indicate the contrast between the Buzet dialect and insular South Čakavian (especially their cakavian varieties) in the first dimension, and between Molise and all cakavian varieties in the Eastern Adriatic region in the second.

While Figure 8a is relevant in that it visualizes the similarity of different varieties in three dimensions, thus accounting for 83.2% of variance in the data (r=0.91), one-dimensional maps (Fig. 8c) are a useful tool to assess the relative importance of different groupings so that the most conspicuous linguistic differences on average are mapped in the first dimension, somewhat less conspicuous ones in the second, and so forth. In the analysis of pronunciation, the difference between the Buzet dialect (esp. Nugla and Sv. Martin) varieties and the cakavian varieties in the South Čakavian region (Hvar, Milna, and the island of Vis) is mapped in the first dimension (r=0.59). Here a few additional features set South cakavian varieties apart from North Istria such as the ikavian reflex of jat, the closing of a /a/ > ạ /ɑ/ or ọ /ɔ/, as well as the absence of syllabic r /ṛ/ and word

final devoicing and reductions in the speech of Hvar, Brač and Vis. The difference between Molise Croatian and all cakavian varieties on the eastern Adriatic coast in the second (r=0.66) is the most pronounced one in the second dimension, which is due not only to cakavism, but also to different accentual patterns are responsible for the mentioned contrast (e.g. /ʑɛnâ/ 'woman' and /ɕɛɕtrâ/ 'sister' in all cakavian varieties vs. /ʒˈɛːna/ and /sˈɛːstra/ in Molise). The mapping of the cakavian varieties in the first three dimensions (in the third dimension r=0.46) is due to the specific realizations of sibilants as well as the absence of the postalveolar affricate.

4.2. Analysis of lexical differences

Lexical distances were calculated by means of a categorical analysis of the values of the differing cognate classes assigned to the lexical variants of the 92 concepts (Cronbach α=0.84). If the lexemes noted are not local or idiolectal expressions, this kind of analysis could point to linguistic influences which may also correspond to historical developments, just as the differences at the phonological level do. Linguistically, we expect the notoriously volatile lexicon to reflect the influences of recent history more readily than phonology does.

As pointed out in section 3.1, 84 concepts were attested in at least 85% of locations. The root was the same in all 88 varieties for 44 of the 84 concepts, while for other 40 we encountered different roots as the basis of the relevant lexeme. For those 40 concepts a root different from the predominant one was found in a small number of varieties (ranging from one to up to thirteen. For the remaining eight concepts (in bold in Table 2) we found no predominant root. In these cases we found either a proliferation of roots used for a single concept (up to 10 for words such as *prišt* 'pimple' and *cvijet* 'flower') or a small number of cognates equally split among the varieties (e.g. *djed – nono* 'grandfather').

Apart from the varieties spoken in Molise where the influence of contact with Italian varieties is much more obvious on the lexical level than at the level of pronunciation, the influence of the Romance superstrate is not felt equally in all the varieties along the Adriatic coast. In many cases the varieties also differ on the basis of

the retention of a specific Slavic root, e.g. where more than one is found in the analyzed vocabulary (e.g. *kiša* 'rain' has been recorded only in the Štokavian varieties on the island of Pag, Pelješac and in Makarsko primorje, while in all others *dažd* < PS *dъždь is used).

4.2.1. Cluster analysis

Compared to clustering based on pronunciation differences, the probabilistic clustering based on lexical differences is characterized by important higher level groupings. One colleague suggested that this might be a consequence of analyzing categorical information, which is less sensitive and might therefore be more easily grouped. But categorical data has no inherent tendency to lend itself more readily to clustering. The larger clusters found in the lexical grouping could, however, indicate an important role played by language contact, which may have caused the varieties to converge more on the lexical than on the phonological level.

Lexical clustering also indicates the divergence of Molise Croatian, undoubtedly due to lexical borrowings from the surrounding Italian dialects with which it has been in contact since the last important migratory wave from the overseas homeland in the 17th century (Sujoldžić, 1990). All other varieties form a single cluster, which is split into two parts at a lower level: Istria on the one hand and the rest of the eastern Adriatic varieties on the other (Fig. 9a). Although Istria lexically forms a cluster (98% certainty), the internal divisions follow approximately the one based on pronunciation so that southwestern ikavian varieties, Central Istrian North Čakavian, and the Buzet dialect each form separate clusters. North Čakavian varieties in Istria do not form a uniform cluster, but rather split into several smaller clusters and a few outliers. All other varieties outside Istria form one large stable cluster (84%). Within that cluster, however, the clustering again is consistent along geographic lines, although a separate cluster including all the Štokavian ikavian varieties (or Čakavian with significant Štokavian influence) is more prominent here than in the analysis based on phonetic and prosodic differences. This cluster includes the coastal varieties of Krilo Jesenice, Omiš and Zaostrog; Bogomolje, Svirče, and Sućuraj on Hvar; Sumartin and Novo Selo on Brač (Fig. 9a & 9b).

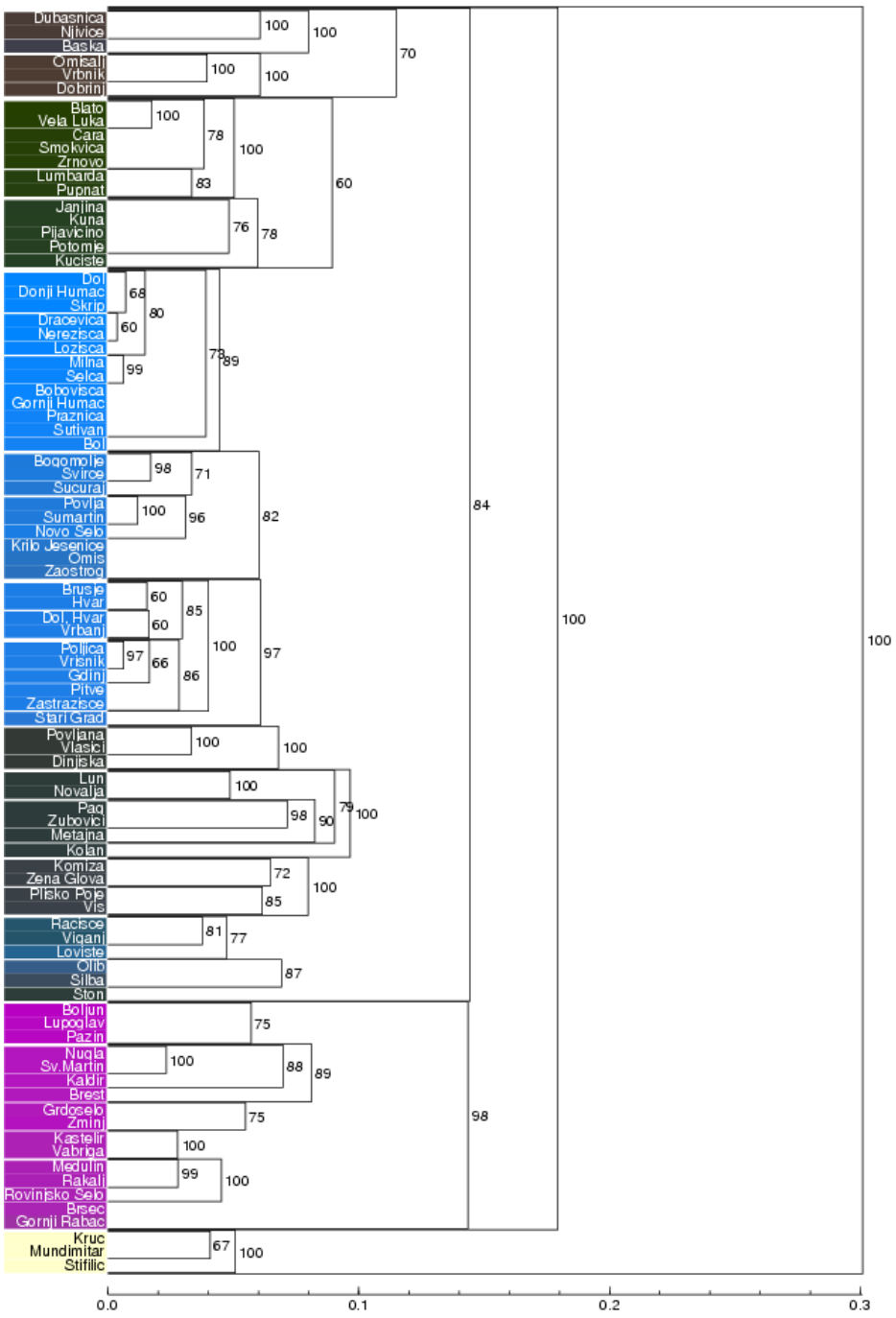

Figure 9a. Probabilistic clustering based on lexical differences only. The x-axis is the mean normalized Levenshtein distance between varieties (noise=0.2sd, group and weighted average combined).

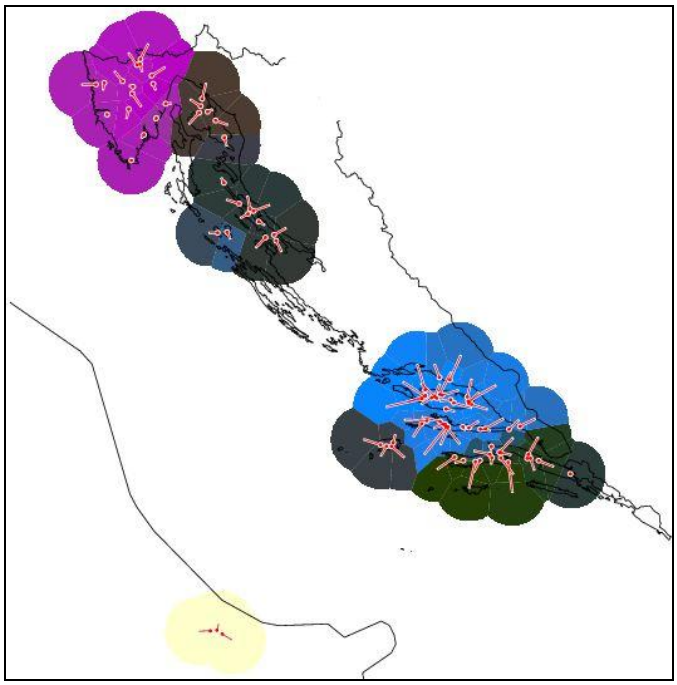

Figure 9b. Probabilistic clustering of lexical differences enhanced visually by multidimensional scaling shows the major cluster divisions based on the combined group and weighted average clustering (noise=0.2sd).

4.2.2. Multidimensional scaling

Although MDS often provides a representation of gradual transitions, it is clear that in the analysis of lexical data, Molise visibly stands apart from all other varieties (Fig. 10a), and indeed more saliently here, where we view the varieties lexically, than it did earlier when they were compared on the basis of pronunciation. In fact, Molise Croatian is mapped in the first two dimensions: against North Čakavian in Istria in the first dimension (r=0.81), and against the whole eastern Adriatic coast in the second (r=0.76) (Fig. 10b). In both cases this is the consequence of a high degree of Romance influence, which is manifested more on the lexical than on the phonological level. Although it might seem perplexing at first that the Romance lexical influence sets Molise Croatian apart so much from other eastern Adriatic varieties if those were exposed to centuries-long direct contact with Venetian and indirectly with the Italian literary language, the Romance-based loanwords differ both in quantity and kind in the two groups due to the patterns and intensity of

linguistic contact, as well as the Romance varieties they came into contact with. Another factor contributing to such a lexical distance between the two groups is due to lexemes not found elsewhere (e.g. *'tarela'* grandfather) and unusual semantic and/or morphological transfers (e.g. *ljud* 'man') in Molise Croatian. Istrian North Čakavian, especially its northern Buzet dialect not only differs from Molise Croatian, but also remains distinct from the rest of the coastal varieties due to a higher proportion of Romance loanwords (e.g. *pištrin* 'grindstone', *kunfin* 'boudary stone', etc.), and a number of Slavic words not present in the more southern varieties, some of which are preserved in the northern South Slavic varieties (e.g. *brek / brak* 'dog', *perje* 'leaves', *otrok* 'child', etc.) Krk is lexically more similar to Istria than to the rest of the varieties in this representation, while the whole island of Pag together with Silba and Olib forms a continuum with the varieties in the south.

Orange hues visible in the south on the MDS map (Fig. 10a) separate the varieties of Korčula and western Pelješac from the rest of South Čakavian varieties, but also from neighboring ikavian and jekavian Štokavian varieties in the Makarska region and in the eastern part of Pelješac. These varieties, in which the Čakavian substrate has been exposed to a Štokavian superstrate, are distinguished in the third dimension accounting for 14.4% of the total variance (in the original lexical differences). A point of interest might be the fact that they contrast with all four varieties on the island of Vis, but do not contrast with other South Čakavian varieties.

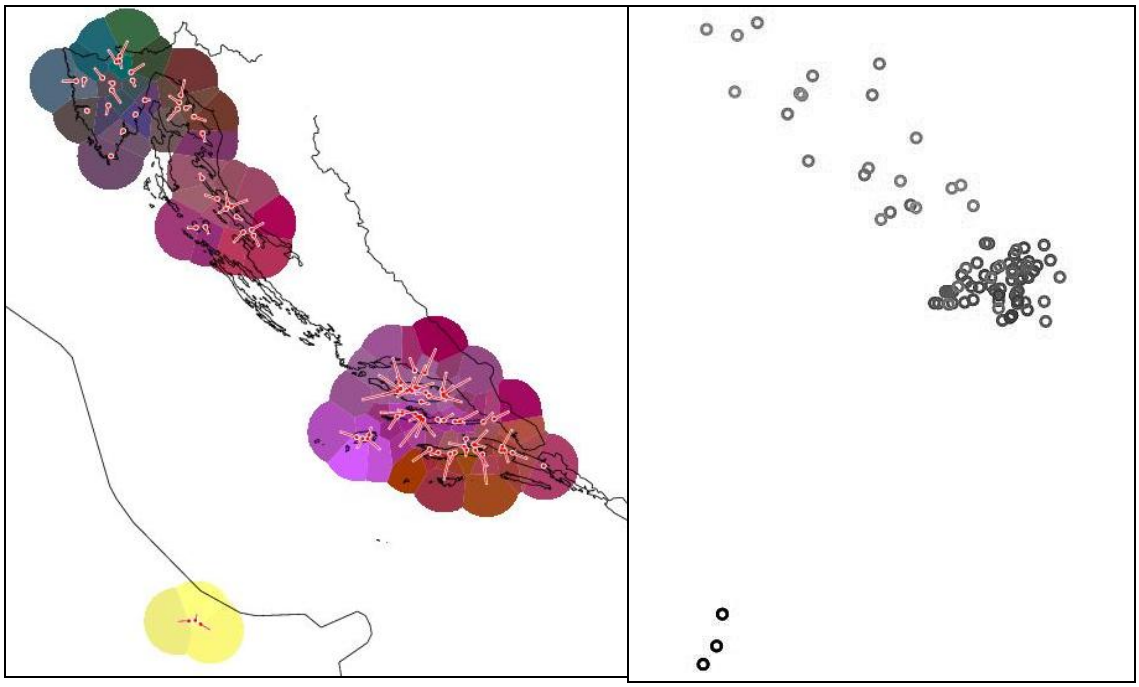

Figure 10a. Lexical differences represented by MDS map in RGB three-dimensional space (r=0.94) and a MDS plot (r=0.91).

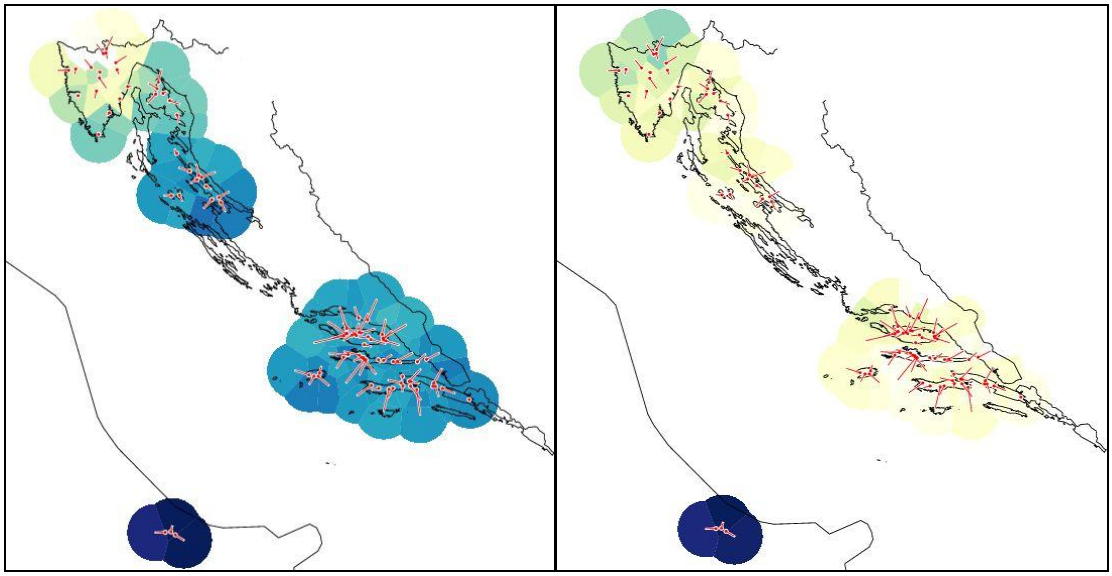

Figure 10b. One-dimensional lexical MDS maps indicate the contrast between Molise and north Istria in the first dimension, and between Molise and most varieties in the Eastern Adriatic region in the second.

In this lexically based analysis (Fig. 10a) there are no 'language islands' that significantly disrupt the continuum the way the cakavian varieties appeared as islands in

the MDS analysis based on pronunciation (Fig. 8a). Another difference from the earlier pronunciation analysis is that the varieties spoken in SW Istria lexically do not show the same conspicuous similarity to the Štokavian (or Štokavian-influenced) varieties on Pag, Makarska coast, Korčula, and Pelješac that we encountered in the analysis based on pronunciation. Both examples could be indicative of a high degree of linguistic contact that cannot be discerned on the basis of pronunciation analysis alone, assuming that pronunciation reflects contact influences less immediately.

5. Conclusions and prospects

The pronunciation analyses have shown that the varieties investigated along the Adriatic coast form neither easily distinguishable dialect areas, nor linguistic continua; the whole region is marked rather by discontinuity made up of many small clusters and no true transitional zones (cf. Brozović, 1960, 1970). This is, moreover, more true of the northern Adriatic area than of the southern. The greater linguistic distances in the north are in accordance with earlier dialectological scholarship according to which there are more distinct dialects in that area, especially in Istria. Such diatopic diversity can be attributed to the numerous migrations of the Štokavian ikavian speakers from the south mostly between the 15th and 17th centuries (Brozović, 1970; Kalsbeek, 1998:24; Małecki, 2007: 158-159; Lisac, 2009). When migrants came in groups with strong social networks, the settlements they founded and populated remained demographically, culturally and linguistically rather homogeneous and distinct from their new neighbors. Although a certain amount of lexical leveling occurred due to contact with Romance dialects (primarily Venetian), the leveling was not nearly as pervasive at the phonological level as at the lexical level. This at least partly explains the linguistic divisions found in the northern Adriatic region, particularly in Istria.

In contrast to the situation in the north, the analysis has shown that the pronunciation and lexical divisions between Štokavian and Čakavian varieties in the south are not nearly as large. There the geographic lines, and more specifically island shores, seem to enhance linguistic cohesiveness in both pronunciation and vocabulary. This supports some of the earlier findings of lexicostatistical investigation of linguistic variation in Middle Dalmatia (Sujoldžić, 1997:296). In most cases the grouping of the

varieties in individual micro-regions (islands, peninsulas) follows the patterns noted in previous studies, and occasional differences can be ascribed to different statistical approaches in handling the lexical data (see 1.2).

One of these differences regards the position of the Molise Croatian dialect. In both Sujoldžić 1990 and Sujoldžić 1997, the Croatian varieties spoken in Molise were grouped with Štokavian or Čakavian-Štokavian dialects in the southern Adriatic, while they form a decidedly separate cluster (Fig. 7a) in the present study. Another difference concerns the make-up of the cakavian cluster, which comprised more varieties in this study based on the Levenshtein distances compared to an earlier one (Sujoldžić, 1997). The third important difference derives from the amount of mixing of various adstrates on the island of Korčula, which is reflected in its very unstable position in clustering with respect to the rest of the varieties investigated. While in earlier lexicostatistical analysis the whole island (except Račišće) was grouped with other Čakavian varieties (Sujoldžić, 1997), in the present study it formed a cluster with other Čakavian-Štokavian varieties, and in the analysis of both pronunciation and vocabulary it formed a group with Pelješac (Fig. 7a & 9a). The dialectally transitional character of Korčula has been best depicted by the application of multidimensional scaling (Fig.8a).

Based on the present analysis of a relatively large area it can be concluded that the internal diversification of what is normally referred to as Čakavsko *narječje* is considerable. This applies not only to the comparison of North and South Čakavian varieties, but also to the comparison between neighboring cakavian and Čakavian speech habits. The problem encountered in the attempt to group varieties on the island of Korčula as predominantly either Čakavian or Štokavian thus points to the need to: a) reflect on the current practice of insisting on assigning all varieties to clearly delineated and non-overlapping (groups of) dialects, and b) critically approach the conclusions based on clustering methods as they may conceal aspects of the linguistic reality, which is seldom as clear-cut as suggested by (hard) clustering.

The results of the present study also call for reconsideration of some widely-accepted dialectal classifications in Croatian dialectology based on the selection of isoglosses. It is true that diachronically informed approaches tend to group dialects on the basis of shared innovations and thus to disregard the similarities based on the preservation of archaic features. From the synchronic perspective, which cannot be

neglected altogether when discussing geolinguistic variation, focusing only on a few carefully chosen features – regardless of how relevant they might be from a historical perspective – ignores similarities between the varieties not sharing a certain isogloss or an innovation. Because the presence or absence of features determining isoglosses are sometimes found only rarely (in few and/or infrequent words), they do not always reflect realistic distances between varieties as dialect speakers perceive them. Measuring and aggregating linguistic distances takes into account both the similarities and differences found in different varieties. As in other studies based on purely quantitative methodology, a number of conditions have to be met concerning the choice and amount of data, the selection of the respondents, transcription quality and the normalization of the transcriptions (when relevant), and finally the methodological choices taken in statistical analysis (tokenization, weighing of features if applicable, etc.).

After calculating both pronunciation and lexical distances, we correlated the two in order to check the extent to which the analyses agree. The differences in MDS representations indicated that the two kinds of variation differ considerably at least in some respects, we could not be certain how closely the two sorts of variation jibed with each other. The correlation between phonological and lexical distances is $r=0.72$ ($p<0.000001$), which is quite substantial, but which explains (only) 50% of the variance in the data. So on the one hand the two sorts of variation probably do reflect similar dynamics, presumably those of close contact, but on the other hand they differ significantly as well. We conjecture that similarities in phonological and prosodic features tend to reflect historical (genetic) relationships more faithfully than lexical similarities, which in turn reflect the effect of contact more truly. In the lexical analysis more gradual transitions were observed between different areas compared to the often scattered groupings obtained on the basis of pronunciation analysis (for example, in lexically-based analysis we found nothing resembling the strong but geographically dispersed cakavian cluster detected in the analysis based on pronunciation).

It is clear that the inclusion of both pronunciation and lexical data in the analysis contributed to better understanding of dialectal diversity found in the region. It also points to the need to extend the analysis by including morphological and possibly syntactic levels in subsequent studies as all of them have an effect on linguistic differentiation as well as on mutual intelligibility, which in turn further promotes

dialectal convergence and/or divergence. We also hope that a wider and more balanced coverage of a larger geographic area as well as an increase in the number of items collected might supplement the present database in order to contribute to a more reliable account of dialectal diversity in the region.

Acknowledgements

The authors would like to thank Peter Kleiweg for creating a very user-friendly platform for doing dialectometry and for his help in resolving software-related problems. Rinke Colen wrote programs for data conversion and formatting, and Jelena Prokić helped us with map drawing. We also very much appreciate the generous help and advice concerning the proper method of IPA transcription of the Croatian dialectological data provided by Dunja Brozović Rončević and Ivana Kurtović from the Institute for the Croatian Language and Linguistics in Zagreb. The research on Croatian varieties was funded by the Ministry of Science, Education and Sport of the Republic of Croatia under grants 0196002 and 196-1962766-2743. The collaboration between the Institute for Anthropological Research in Zagreb and the Center for Language and Cognition in Groningen was supported by ENC Coimbra Group Hospitality Scheme and financed by the University of Groningen.

6. References

Bolognesi, Roberto and Wilbert Heeringa. (2005) "De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten". Gramma/TTT: tijdschrift voor taalwetenschap 9(1): 45-84.

Brozović, Dalibor. (1960) "O strukturalnim i genetskim kriterijima u klasifikaciji hrvatskosrpskih dijalekata". Zbornik za filologiju i lingvistiku 3: 68-88.

Brozović, Dalibor. (1970) "Dijalekatska slika hrvatskosrpskoga jezičnog prostora". Radovi: razdio lingvističko-filološki 8: 5-32.

Chambers, J. K. and Peter Trudgill. (1998) Dialectology. Cambridge: Cambridge University Press.

Goebl, Hans. (1984) Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3rd vol. Tübingen: Max Niemeyer.

Goebl, Hans. (2006) “Recent Advances in Salzburg Dialectometry”. Literary and Linguistic Computing 21: 411-435.

Gooskens, Charlotte and Wilbert Heeringa (2004). “Perceptive Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data”. Language Variation and Change 16(3): 189-207.

Gooskens, Charlotte et al. (2010). “Phonetic and Lexical Predictors of Intelligibility”. International Journal of Humanities and Arts Computing 2(1/2): 63-81.

Gusfield, Dan (1997) Algorithms on Strings, Trees and Sequences. Cambridge: Cambridge University Press.

Heeringa, Wilbert. (2004) Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. dissertation, U. of Groningen.

Heeringa, Wilbert and John Nerbonne (1999). “Change, Convergence and Divergence among Dutch and Frisian”. Philologia Frisica: Lêzingen fan it fyftjinde Frysk filologekongres. (Fryske Akademy, Ljouwert), 88-109.

Ivić, Pavle. (1981) “Prilog karakterizaciji pojedinih grupa čakavskih govora”. Hrvatski dijalektološki zbornik 5: 67-91.

Jakubinskij, Lav. (1925) “Die Vertretung des urslav. ě im Čakavischen”. Zeitschrift für slavische Philologie 1: 381-396

Kalsbeek, Janneke. (1998) The Čakavian Dialect of Orbanići near Žminj in Istria. Amsterdam – Atlanta: Rodopi.

Kapović, Mate. (2004) “Jezični utjecaj velikih gradova”. Rasprave Instituta za hrvatski jezik i jezikoslovlje 30, 97-105.

Kessler, Brett. (1995) “Computational dialectology in Irish Gaelic”. Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL, Dublin), 60-67.

Kruskal, Joseph (1999) “An Overview of Sequence Comparison” David Sankoff and Joseph Kruskal (eds.) Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison. Introduction by John Nerbonne. Stanford: CSLI Publications. 1-44.

Landau, Ernestina et al. (1999). “Croatian”. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge: Cambridge University Press, 66-69.

Lisac, Josip. (2003) Hrvatska dijalektologija 1: Hrvatski dijalekti i govori štokavskog narječja i hrvatski govori torlačkog narječja. Zagreb: Golden Marketing – Tehnička knjiga.

Lisac, Josip. (2009) Hrvatska dijalektologija 2: Čakavsko narječje. Zagreb: Golden Marketing – Tehnička knjiga.

Lupić, Ivan. (2001) "Preskriptivna akcentologija i hrvatski standardni jezik". Kolo: časopis Matice hrvatske XI (1), 85-134.

Małecki, Mieczysław. (2007) Čakavske studije. Rijeka: Maveda.

Moguš, Milan. (1977) Čakavsko narječje: Fonologija. Zagreb: Školska knjiga.

Nerbonne, John. (2009) "Data-Driven Dialectology". Language and Linguistics Compass 3(1): 175-198. DOI: 10.1111/j.1749-818x2008.00114.x

Nerbonne, John. (2010) "Measuring the Diffusion of Linguistic Change". Philosophical Transactions of the Royal Society: Biological Sciences 365: 3821-3828. DOI: 10.1098/rstb.2010.0048

Nerbonne, John. (2011) "Mapping Aggregate Variation". Alfred Lameli et al., eds. Language and Space, Vol.2: Mapping Language. Berlin: Mouton De Gruyter, 476-494. DOI: 10.1515/9783110219166.1.fm

Nerbonne, John et al. (1996). Phonetic Distance between Dutch Dialects. In: Durieux, G., Daelemans, W. and S.Gillis (eds.) CLIN VI: Proceedings of the Sixth CLIN Meeting. Antwerp: Centre for Dutch Language and Speech (UIA). 185-202.

Nerbonne, John and Christine Siedle. (2005) "Dialektklassifikation auf der Grundlage Aggregierter Ausspracheunterschiede". Zeitschrift für Dialektologie und Linguistik 72(2): 129-147.

Nerbonne, John et al. (2008). "Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering". Christine Preisach et al., eds. Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society. Berlin: Springer, 647-654.

Nerbonne, John and Wilbert Heeringa. (2010) "Measuring Dialect Differences". Peter Auer and Jürgen E. Schmidt, eds. Language and Space, Vol. 1: Theories and Methods. Berlin: Mouton De Gruyter, 550-567. DOI: 10.1515/9783110220278

Nerbonne, John et al. (2010) "Applying Language Technology to Detect Shift Effects". Muriel Norde et al., eds. Language Contact: New Perspectives. Amsterdam: Benjamins, 27-44.

Nerbonne, John et al. (2011). “Gabmap – A Web Application for Dialectology”. Dialectologia. Special Issue II, 65-89.

Prokić, Jelena and John Nerbonne. (2008) “Recognizing groups among dialects”. International Journal of Humanities and Arts Computing 2: 153-171.

Prokić, Jelena et al. (2009) “The Computational Analysis of Bulgarian Dialect Pronunciation”. Serdica Journal of Computing 3(3): 269-298.

Séguy, Jean. (1973) “La dialectométrie dans l'Atlas linguistique de la Gascogne”. Revue de Linguistique Romane 37: 1-24.

Spruit, Marco René et al. (2009) “Associations among Linguistic Levels”. Lingua 119(11): 1624-1642.

Srce & Novi Liber. Hrvatski jezični portal (HJP). <http://hjp.srce.hr> [last accessed: 05/2011]

Sujoldžić, Anita. (1989) “Cultural Microevolution of the Islands of Silba and Olib Measured by Linguistic Data”. Collegium Antropologicum 13: 189-195.

Sujoldžić, Anita. (1990) “The analysis of population histroy and cultural (linguistic) microevolution of the Slavic settlements in Molise, Italy”. HOMO 41: 1-15.

Sujoldžić, Anita. (1994) “Govori srednjodalmatinskog otočja: prilog antropologijskim istraživanjima”. Društvena istraživanja 12/13: 423-436.

Sujoldžić, Anita. (1997) “Continuity and Change Reflected in Synchronic and Diachronic Linguistic Variation of Middle Dalmatia”. Collegium Antropologicum 21: 285-299.

Sujoldžić, Anita et al. (1982/83) “Lingvističke udaljenosti na otoku Hvaru”. Rasprave Zavoda za jezik 8/9: 197-214.

Sujoldžić, Anita et al. (1986). “Linguistic Microdifferentiation on the Island of Korčula”. Anthropological Linguistics 28: 405-432.

Sujoldžić, Anita et al. (1988). “Sličnosti i razlike u govorima otoka Brača kao odraz migracijskih kretanja”. Rasprave Zavoda za jezik IFF 14: 163-184.

Sujoldžić, Anita et al. (1989) “Jezične udaljenosti na poluotoku Pelješcu”. Ljubljana: SAZU.

Sujoldžić, Anita et al. (1990) “Lingvističke udaljenosti otoka Paga”. Filologija 18: 7-37.

Sujoldžić, Anita et al. (1992/93). “Govori otoka Krka - Uvod u antropološka istraživanja”. Filologija 20/21: 431-449.

Swadesh, Morris. (1952). “Lexicostatistic dating of prehistoric ethnic contacts”. Proceedings of American Philosophical Society 96: 452–463.

Szirovicza, Lajos et al. (1997). “The Comparison of Two Methods in the Anthropological Study of Linguistic Differentiation”. Collegium Antropologicum 21: 609-619.

Šimičić, Lucija. (2005). Analiza temeljnog i kulturnog leksika viških govora i njihov odnos prema drugim istraživanim govorima istočnog Jadrana. Pre-doctoral thesis (unpublished), U. of Zagreb.

Šimičić, Lucija and Anita Sujoldžić. (2009). “Istraživanje temeljnog i kulturnog leksika u naseljima otoka Visa – prostorni i generacijski aspekti”. Ines Prica and Željka Jelavić, eds. Destinacije čežnje, lokacije samoće – uvidi u kulturu i razvojne mogućnosti hrvatskih otoka. Zagreb: HED, 189-202.

Škreblin, Lana et al. “Ethnohistorical Processes, Demographic Structure, and Linguistic Determinants of the Island of Vis”. Collegium Antropologicum 26 (2002): 333-350.

Thomason, Sarah and Kaufmann, Terence. (1988) “Language contact, creolization, and genetic linguistics”. Berkeley: University of California Press.

Valls, Esteve et al. (in press). “Applying Levenshtein Distance to Catalan Dialects. A Brief Comparison of Two Dialectometric Approaches”. Verba: Anuario Galego de Filoloxía.

Vermeer, Willem. (1982) “On the principal sources for the study of čakavian dialects with neocircumflex in adjectives and e-presents”. Studies in Slavic and General Linguistics 2: 279-341.

Wieling, Martijn et al. (2009) “Evaluating the Pairwise String Alignments of Pronunciations”. Lars Borin and Piroska Lendvai, eds. Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities and Education (LaTeCH - SHELT&R). Athens: Association for Computational Linguistics, 26-34.