Degrees of non-standardness. Feature-based analysis of variation in a Torlak dialect corpus

Vuković, Teodora ; Escher, Anastasia ; Sonnenhauser, Barbara

DOI: https://doi.org/10.1075/ijcl.20014.vuk

Degrees of non-standardness

Feature-based analysis of variation in a Torlak dialect corpus

Teodora Vuković, Anastasia Escher, and Barbara Sonnenhauser[1]

Slavisches Seminar, University of Zurich

Published in:
Teodora Vuković, Anastasia Escher, Barbara Sonnenhauser (2022). Degrees of non-standardness. Feature-based analysis of variation in a Torlak dialect corpus. *International Journal of Corpus Linguistics*. John Benjamins. DOI: https://doi.org/10.1075/ijcl.20014.vuk

A corpus-based method for assessing a range of dialect-standard variation is presented for identifying samples exhibiting the highest prevalence of dialect features. This method provides insight into areal and inter-speaker variation and allows the extraction of maximally non-standard manifestations of the dialect, which may then be sampled and used for the study of language change and variation. The focus is on a non-standard Torlak variety, which has undergone considerable change under the influence of standard Serbian. The degree of variation is assessed by measuring the frequencies of five distinguishing linguistic features: accent position, dative reflexive *si*, auxiliary omission in the compound perfect, the post-positive article, and analytic case marking in the indirect object and possessive. Locations subject to the greatest and least influence of the standard are revealed using hierarchical clustering. A positive correlation between the frequencies of occurrence reveals which non-standard feature is the best predictor of the others.

Keywords: Linguistic variation, corpus-based dialectometry, endangered languages, spoken language, Torlak

1. Introduction

[1] The authors' responsibilities were as follows. Editing: TV; choice of features: TV, AE, BS; corpus analysis for accent position, post-positive demonstratives, *si* particle: TV; corpus analysis for AUX omission: AE, TV; statistical analyses: TV, AE; linguistic embedding into the South Slavic context; linguistic variation expertise, mentoring: BS.

The South Slavic dialect continuum is characterized by an intricate encounter of affiliations. Genealogically, it is intersected by an old bundle of isoglosses differentiating Western and Eastern South Slavic. From the areal point of view, parts of it share a number of morphosyntactic innovations with their neighboring non- Slavic languages. That is, while the contemporary standard languages of Serbian, and Macedonian, and Bulgarian belong genetically to the South Slavic branch, their noun phrase and verbal systems diverge in crucial aspects from one another. Macedonian and Bulgarian, but not Serbian, exhibit traits characteristic of the Balkan linguistic area (see, e.g. Lindstedt, 2000, for a detailed description of Balkan morphosyntactic features). The resulting variation becomes most distinct in the Torlak dialects spoken in Southern Serbia.

2. Variation in Torlak

By their archaic features, mainly phonological and morphological, the Torlak dialects fit in with Serbian in the western range of the South Slavic dialect continuum. At the same time, they differ from Western South Slavic by not having undergone the Neo-Shtokavian accent shift otherwise characteristic of Serbian dialects and standard Serbian. The innovative, mostly morphosyntactic, aspects of Torlak are characteristic of both the Eastern South Slavic group, like Bulgarian and Macedonian (Ivić, 2009:152; Hinrichs, 1999), and the Balkan linguistic area. Another factor is the present vertical influence of Standard Serbian. These mixed affiliations make linguistic descriptions of Torlak challenging, especially when the aim is to capture the variation between the dialect's most conservative / non-standard / East Slavic manifestation, and change towards the West Slavic, accelerated by the influence of the standard. A distinction between the dialect and the standard can be made by a set of representative linguistic features. The use of the dialectal variant of the features is viewed as a baseline against which variation towards the standard can be assessed. Since Torlak is an endangered language, identifying and describing the variation is a crucial tool for language documentation. In addition, it contributes to a better understanding of dialectal and areal feature diffusion in a diachronic and diastratic context.

2.1 Dimensions of variation

Approximating the Torlak base dialect necessitates the disentangling of variation encountered in the speech of individual speakers. In the diastratic dimension, this pertains to the variation

between dialect and standard, i.e. between Torlak and standard Serbian. While this kind of variation results mainly from vertical – or register-based – contact, variation along the diatopic dimension results primarily from horizontal contacts, in this case, contact among Serbian dialects (Old-Shtokavian Torlak and Neo-Shtokaian Serbian), and contact across Eastern and Western South Slavic (South Slavic Torlak with Balkan Slavic Bulgarian and Macedonian) dialects. The diachronic dimension includes the variation of archaic and innovative features. Elaborating on these dimensions of variation thus emerges as a necessary prerequisite for a proper description of Torlak. At the same time, it contributes to placing dialectology within areal-typology by assessing the interaction of innovations diffused through language contacts with the inherited genealogical features.

Disentangling variation is of particular relevance when it comes to documenting a dialect for both its preservation as cultural heritage as well as its linguistic description. Experience from fieldwork shows that individual criteria, be they social, demographic, or geographic,[2] cannot be used as reliable predictors for the degree of influence from the standard variety. Their separate impact or interplay cannot always be discerned on a general scale. In the case of Torlak, assessing variation is further complicated by its complex sociolinguistic embedding. Anthropological studies have indicated that these dialects are stigmatized (Krstić, 2014:24–153)[3] by speakers from other dialect regions which regard them as indicators of a low level of culture (Vuković & Samardžić, 2018: 185; Petrović, 2015). As a consequence, Torlak speakers try to disguise their dialect as much as possible and avoid using it in communication with outsiders, making data hard to obtain. It is generally known that salient features tend to disappear first in a dialect in contact with a prestigious variety (Trudgill, 1986: 37). Presently, the increasing availability of education, the growing influence of the standard variety, and the demise of the elderly population all contribute to turning dialectology into linguistic archaeology. For instance, the dwindling of this dialect (on the typology of these processes, see Chambers and Trugdill 1998: 69–76) can be seen by the more intense standard Serbian influence samples dating to the beginning of the 21st century (Sobolev, 1998) as compared to samples collected at the beginning of the 20th century (Belić, 1905; Stanojević, 1911).

[2] Concerning postposed definiteness marking, Vuković and Samardžić (2018) show that speakers in remote villages in high altitudes use this feature more frequently than other speakers. However, this finding is difficult to generalize.

[3] The word *Torlak* can even be used pejoratively in Serbian, meaning "bull-headed" or "stupid" (Krstić, 2014: 571).

This intricate situation requires strategies of data sampling and tools of data processing that enable the assessment of the degrees of variation and, hence, the degree of "Torlakness" in single samples. In order to avoid making claims based on potentially misleading extra-linguistic factors and intuitive observations (e.g. most younger speakers tend to use a more standardized variety of the dialect, but not always), variation needs to be explored based on linguistic features alone.

2.2 Assessing variation

The first step in describing Torlak is finding a means of identifying it. This in turn presupposes baselines and points of reference for establishing criteria for differentiating Torlak from Neo-Shtokavian dialectal and standard Serbian, and defining a documented historical starting point in order to assess feature diffusion and language change.

The documentation of Torlak from the late 19th/early 20th centuries by Belić (1905) and Stanojević (1911) is taken as a historical baseline. These data provide examples of the language of older people at the beginning of the 20th century. Some of them had been born before the Serbian standard was established in its modern form and many of them had received no formal education. Since this essentially means that these speakers had little or no contact with standard Serbian throughout their lives, this variety documents Torlak in its most authentic manifestation. Towards the end of the 20th century, it becames harder to find speakers who had remained largely uninfluenced by the standard variety, although a few are documented in Sobolev (1998). Thus, the samples provided by Belić (1905), Stanojević (1911), and Sobolev (1998) can be taken as 'prototypical dialect' in the sense defined above. These samples serve as a baseline for the synchronic estimation of "dialectness" instantiated in the single samples gathered in current fieldwork, as well as for the diachronic tracing of specific features.

Linguistic situations in which a normative standard variety coexists with local dialects are common. In such situations, every single variety and even every single utterance assumes a certain position on a range between the standard variety and a dialect that has been uninfluenced by it. The exact position on the scale depends on the presence of features characteristic of one dialect or a standard variety. In Timok, the distinction between the group of younger and older speakers can easily be made by native speakers, and empirically demonstrated (see Appendix 1, see Vuković et al., forthcoming), but age is not a reliable factor in the older group of speakers who are important for the description of the dialect. While

intuitive classification based on the Torlak and standard Serbian characteristics can provide a very rough and intuitive differentiation between individual samples, such comparisons are often subjective and limited and do not allow for finer-grained assessments and systematic generalizations, especially when the researcher is faced with larger datasets.

In the present paper, a quantitative method for the empirical evaluation of the influence of the standard variety is proposed. It is based on frequency measures of non-standard features and enables the identification of more or less dialectal speech samples. The feature-based approach using corpora has already been established and tested many times in the field of linguistic variation and change, as well as dialectometry (e.g. Nerbonne & Kretzschmar, 2012; Szmrecsanyi, 2017, 2015; Szmrecsanyi & Wälchli, 2014). Apart from providing information about the variation itself, the samples resulting from such an approach can be compared, and the most dialectally characteristic of the speakers in different locations across the region identified. This, in turn, provides a basis for further identifying possible socio-cultural and geographical factors influencing the degree of variation and triggering feature diffusion from the standard variety into the dialect.

3. Torlak features chosen for analysis

As a highly stigmatized dialect, some salient features of Torlak tend to be generally interpreted as shibboleths. These include postponed definiteness marking, the loss of cases, or clitic doubling (e.g. Krstić, 2014) and are often avoided by speakers in interview situations. Other features that distinguish Torlak from Serbian, such as the place of accentuation, the dative reflexive clitic *si*, the omission of the auxiliary or complementizers, are less immediately noticeable and hence are less likely to be consciously avoided and to trigger code-switching. This section describes the features selected for analyzing the variation across speakers, the rationale behind their selection, and the ways they are operationalized in the quantitative analysis of variation. The illustrative examples are taken from the samples from Belić (1905), Stanojević (1911), and Sobolev (1998), containing earlier and minimally influenced attestations of the dialect.

3.1 Selection

The analysis reported in this paper is based on linguistic features that belong to the phonological, nominal and verbal domain:

(i) Accentuation: preservation of the inherited place of the accent in Torlak vs. "Neo-Shtokavian" accent shift in the dialects forming the basis of standard Serbian;
(ii) Short form of dative reflexive *si*: grammatical functions within VP and NP in Torlak which are not used in standard Serbian;
(iii) Perfect tense morphology: omission of 3rd person auxiliary in Torlak vs. preservation of 3rd person auxiliary in standard Serbian;
(iv) Post-positive article: the use of demonstrative clitics on nominals which are not used in standard Serbian;
(v) Analytic case marking in indirect object and possessive: the *na+oblique* constructions in Torlak vs. synthetic inflectional marking in standard Serbian.

Accentuation is one of the most recognizable traits of these dialects, associated with a rural background and a lack of culture (Petrovic, 2015; Krstić, 2014). The feature, 'omission of 3rd person auxiliary' can be found both in Balkan Slavic (Bulgarian and Macedonian) and in standard Serbian. This feature is more frequent in Torlak than in standard Serbian, but, according to the evidence of our native Serbian informants, it is not perceived as a "borrowed" or "foreign" element. Hence, Serbian standard speakers do not necessarily associate it with Southern dialects, but they tend to characterize its frequent use as a rural or distinctly colloquial trait. The clitic *si* may be perceived by standard speakers as a more dialectal element (Krstić, 2014). Since *si* is present in some northern dialects of the BCMS dialect continuum, it is not judged as something completely "foreign", but rather recognized as "south Serbian" and "rural" and therefore as less prestigious. Analytic marking used instead of synthetic case inflections and post-positive articles are typical features of the Balkan Sprachbund. They are perceived as extremely dialectal and considered as one of Torlak's principal characteristics.

Among the three features, the position of the accent – one of the important distinctive features in the South Slavic dialect continuum – has a special role in the analysis. As research on phonetic accommodation shows, phonological features are those features that speakers are least aware of when switching between the standard variety and the dialect (Brulard & Carr, 2013). That is, speakers do not monitor their accents in the same way as they control lexicon or morphology. This suggests the hypothesis that a higher frequency of non-standard accentuation, i.e. lack of Neo-Shtokavian accent retraction, can be used as a predictor of a

higher frequency of appearance of other non-standard features. It also suggests that accent position could function as a way to establish a baseline for assessing variation.

3.2 Accent position

Accent in Torlak is not affected by the accentual retraction characteristic of Neo-Shtokavian dialects (in various smaller subdialects there is a great deal of variation and accentual isogloss (Alexander, 1975: 517)) and hence also of the standard Serbian language. This makes accent position a suitable feature for the identification of Torlak samples and Torlak passages within a larger dataset, such as is attempted here.

For analysis several frequent lexemes were chosen that have distinctive standard and non-standard variants with different accent positions. In our sample, we searched for each of the accentuated variants. Examples of those lexemes are provided in Table 1.

Table 1. Accent position in Torlak vs. Serbian

Torlak	**Serbian**	
žená	*žèna*	woman.F.G.NOM
ručák	*rúčak*	lunch-M-SG-NOM
deté	*déte*	child.N.SG.NOM
kojí	*kòji*	who.M.NOM
mojá	*mòja*	my.F.NOM
išlí	*ìšli*	go.M.PPART

In identifying the degree of dialect usage for the single text samples and stretches of utterances within these samples, it can be assumed that the higher the frequency of non-standard accentuation, the less the speech is influenced by the standard variety. Therefore, the usage of dialect accentuation should correlate with usage of further dialect features, such that speakers with non-standard accents can be taken as representative speakers of Torlak.

3.3 The clitic *si*

Balkan Slavic *si* is the short, unstressed form of the dative reflexive pronoun *sebe*. While it is used very frequently in Bulgarian and Macedonian (Petrova, 2014; Savova, 2017), its usage in

contemporary standard Serbian is discouraged by prescriptive grammars (for example, it is not mentioned in Stevanović's (1986) grammar); whenever it appears in Serbian texts, editors tend to delete it (Frleta, 2010: 1–5). As a result, speakers of standard Serbian often perceive it as a regionalism associated with non-prestigious Southern Serbian dialects.

The clitic *si* can perform several functions in South Slavic varieties. Most often, it is used in adnominal (internal, Example (1), glossed as POSS.REFL) and predicative (external, glossed as REFL.DAT) possessive constructions; see a Bulgarian example in (3). Here the focus is on the former. Adnominal possessive dative clitics represent a dependent within a noun phrase and are morphologically invariant synonyms of declinable possessive pronouns; see (1)[4].

(1) *múž-a* ***si*** *natera-l-a* *te* *zamé-l*
husband-OBL POSS.REFL force-PERF-F and knead-PERF.M
u *tigáń* *káš-u*
in pot porridge-OBL
"She forced her husband, and he started cooking porridge in the pot"
(Stanojević 1911: 432)

Predicative possessive dative clitics appear only within a verb phrase, although they are not necessarily part of the valency of the verb. In addition, they are associated with the syntactic role of a non-obligatory indirect object with the semantic shade of the possessor's benefit or harm (dativus commodi / incommodi, dativus sympatheticus, see Arsenijević, 2012). Both patterns, adnominal and predicative, can be observed in Macedonian (Mitkovska, 2011) and Bulgarian, see (2) and (3). The use of *si*, as illustrated in Examples (2) and (3), is ungrammatical in standard Serbian.

(2) Bulg. Possessive clitic *si* in adnominal position:
Kade *si* *složi-l* *čanta-ta* *si?*
Where be.2SG.PRS put-PERF bag-DEF POSS.REFL
"Where did you put your bag?" (Ivanova & Gradinarova, 2015: 517)

(3) Bulg. Possessive clitic *si* in argument position:
Kade *si* *si* *čanta-ta* *složi-l?*
Where be.2SG.PRS REFL.DAT bag-DEF put-PERF

[4] Examples are from Timok dialect, unless otherwise stated.

"Where did you put your bag?" (Ivanova & Gradinarova, 2015: 517)

Another widespread function of *si* in Torlak is middle voice marking (also referred to as evaluative (Arsenijević, 2012) or expressive-emotional meaning (Petrova, 2015); see (4)). This function obscures and neutralizes the opposition between the agent and the patient, indicating that the action is being made in the interest of the subject:

(4)	*Oná*	*se*	*rasrdi*	*te*	***si***
	She	REFL	get angry. 3SG.AOR	and	REFL.DAT
	otide	*u*	*ńón-u*	*sób-u*	
	leave.3SG.AOR	in	her-OBL	room-OBL	

"She got angry and went to her room" (Belić, 1905: 669)

The third possible function of the Torlak *si* is the marking of a prototypical indirect object, a participant in a three-valent verb, e.g., a recipient in (5).

(5)	*doš-l-á*	*bába*	*da*	***si***	*zóvne*
	come-PERF-F	old lady	SUBJ	REFL.DAT	call.3SG.PRS
	zét-a	*i*	*dəšter*		
	son-in-law-OBL	and	daughter		

"The old lady arrived in order to call to herself [=*si*] her son-in-law and her daughter" (Stanojević, 1911: 433).

All three uses of *si* are expected to occur in the data for contemporary Torlak.

3.4 Omission of 3rd person auxiliary with *l*-perfect

The feature, "omission of 3rd person auxiliary with *l*-perfect" differs from the first two features in that it is not perceived as a clear regionalism (dialectism) and hence is not evaluated as strikingly pejorative by the speakers of standard Serbian.

Instances of 3rd person auxiliary omission for the *l*-perfect are not so rare in standard Serbian (the phenomenon was first described in detail in Grickat, 1954; a more recent analysis is provided by Meermann & Sonnenhauser, 2016). This feature is most common in the colloquial register. Omitting the auxiliary carries a certain epistemic semantic load, indicating

the speaker's distancing from the proposition conveyed, e.g. in contexts of conjecture, as in (6).

(6) *Mislim* *znaš* *šta* *to* *znači?*
think.PRS.1SG know.PRS.2SG what this Means
Kad *[-AUX]* *on* *jadnik* *uhvati-o* *Pa*
So he poor man take-PART.PAST.M.SG And
krpio *i* *super* *Iskrpio* -
fix-PART.PAST.M.SG and very well fix-PART.PAST.M.SG
kad *ono* *radi* *kaže* *fala* *Bogu!*
so it work.PRS.3SG say.PRS.3SG thanks God
"Do you know, what I think this means? Then this poor man took and fixed it [the loudspeaker] and he fixed it very well – so it works, he says, thanks God!" (Meerman & Sonnenhauser, 2016: 98).

From a diachronic point of view, the variation of the auxiliary verb can be called a "by-product" of the development of the *l*-forms from a perfect into a general unmarked past tense (e.g. Meermann & Sonnenhauser, 2016: 107). In standard Macedonian, the omission of the 3rd person auxiliary in the analytic perfect tense is mandatory, while in Bulgarian its use or omission can have a semantically distinctive function and serve as a discourse pragmatic marker of changes in perspective (Meerman & Sonnenhauser, 2016: 85–86).

As the potential to omit the 3rd person auxiliary is a regular feature of Balkan Slavic, but largely excluded from the Serbian standard, its occurrence can be taken as a proxy for the distance between Torlak and standard Serbian and the proximity to Balkan Slavic. As a clitic, the auxiliary usually behaves like a Wackernagel element. The past participle can be put in any position in a clause. The clitic can immediately precede it or be separated by other constituents, usually not more than two; see (7) and (8). It may also follow the participle, in which case there are usually no constituents in between, as in (9) (for a more detailed analysis of AUX omission in Torlak, see Escher, 2021, Vuković et al., forthcoming).

(7) [+ AUX] *Neki* *su* *jeli* *ranije* *kukuruznic-u*
some AUX eat-PART.PERF-PL earlier corn-OBL
"Some used to eat corn"

(8) [+AUX] *Žene* *su* *ovako* *sedele* *sa*

	woman-PL	AUX	so	sit-PART.PERF-PL	from
	obe	*strane*			
	both	side-PL			
	"The women were sitting like this from both sides"				
(9) [+AUX]	*Bi-l-i*		*su*	*Nemci*	
	be-PART.PERF-PL		AUX	German-PL	
	"They were Germans"				

3.5 Post-positive article

Post-positive articles are one of the most recognized features of the Balkan Slavic varieties, differentiating them from other Slavic languages. They are standardized in Bulgarian and Macedonian and also used in the Timok variety of Torlak (Joseph, 1992; Belić, 1905; Ivić, 1985). In Torlak, they are one of the most salient dialectal characteristics, and speakers often avoid using them in contact with standard Serbian speakers. Their usage shows considerable inter- and intra-speaker variation. Some speakers, mostly older and living in villages (Vuković & Samardžić, 2018; Vuković, et al., forthcoming), use them relatively frequently, while others do not use them at all.

Etymologically, post-positive articles relate to demonstrative pronouns. Demonstrative pronouns are typically accentuated and precede the noun they modify, as in see (10)a, while post-positive articles are not accentuated and attach to the right of their host, as in (10)b. They both agree with the noun in gender, number and case; see (10)b, (10)c, (10)d. They follow phonetic agreement, based on the ending of the host.

(10) a. *tá* *žená*
that.F.SG.NOM woman.F.SG.NOM
"that woman"

b. *žená-ta*
woman.F.SG.NOM-DEM.F.SG.NOM
"the woman"

c. *ženú-tu*
woman.F.SG.ACC-DEM.F.SG.ACC
"the woman"

d. *žené-te*
woman.F.PL-DEM.F.PL
"the women"

They take the second position in the noun phrase and attach to the left-most element, which can be a noun or modifiers such as adjectives, possessive pronouns or numerals (11).

(11) a. *tá* *žená*

b. *žená-ta*

that.F.SG.NOM woman.F.SG.NOM	woman.F.SG.NOM-DEM.F.SG.NOM
"that woman"	"the woman"

In standard Bulgarian and Macedonian, these post-positive morphemes are fully grammaticalized as definite articles, while in Torlak, where they are often defined as "articles with a strong demonstrative meaning" (Ivić, 1985: 116–117), they appear much less frequently and have not been sufficiently studied.

3.6 Analytic dative marking of the possessive and indirect object

The reduction of the system of inflectional marking on nouns and adjectives is one of the characteristics of Balkan Slavic (Mišeska-Tomić, 2004; Sobolev, 2003; Sobolev, 2008; Wahlström, 2015). Macedonian has completely lost its inflections, coding the direct object with clitic pronominal indexes and their indirect objects (and other 'dative' relations) by means of a pronominal index on the head and prepositional phrasal marking + generalized inflectional marking (na + *casus rectus generalis*) on the dependent (except for some archaic dialects). South-eastern Serbian dialects have kept several relics of inflection marking (casus obliquus generalis) (see Table 2 (Escher, 2021)) used together with prepositional marking.

Table 2: Inventory of inflectional markers of grammatical relations of nouns in Timok[5]

	F		**AM**		**IM**		**N**	
	Sg	**Pl**	**Sg**	**Pl**	**Sg**	**Pl**	**Sg**	**Pl**
Casus rectus	*sestra* "sister"	*sestr*e	*ovčar* "shepherd"	*ovčar*i	*nož* "knife"	*nož*i	*selo* "village"	*sela*
Casus obliquus generalis	*sestr**u***	*=CR*	*ovčar / ovčar**a***	*=CR*	*=CR*	*=CR*	*=CR*	*=CR*

Personal pronouns, on the contrary, due to their high position on the agentivity (or animacy) scale, remain quite resistant to changes of inflectional system; in the Timok variety, the declension of personal pronouns have kept the distinction between accusative and dative, and have lost other peripheral cases.

[5] AM = animate masculine, IM = inanimate masculine, CR = *casus rectus generalis*.

While older sources display the analytic *na* + *accusative* construction quite regularly (see Belić, 1905; Stanojević, 1911; SAOSWB, 1998), in modern sources on the dialect, like the *Spoken Torlak dialect corpus 1.0* (Vuković, 2020, 2021), due to the increasing influence of the standard variety, the *na*-construction is used interchangeably with the dative case; see (13a). When it comes to pronouns, the dative form – see (12b) – is used in alternation with the *na* + accusative construction; see (12c) (Vuković at al., forthcoming).

(12) a) *I* *polako* *u* *sebe* *molitve* *čitam* *bogu*
and slowly in myself.OBL prayer.F.ACC.PL read.1SG.PRES God.M.DAT.SG
"And slowly in myself I read prayers to God."

b) *takoj* *meni* *pričali*
that way I.DAT tell.PPART.M.PL
"(They) were telling me like that."

c) *Dadeš* *ti* *na* *njega* *on* *na* *tebe*
give.2SG.PRES you.NOM on he.OBL.SG he.NOM on you.OBL.SG
"You give to him, he (gives) to you."

The alternation of the inflectional (standard Serbian) and analytic (dialect) strategies is also attested by the expression of possession both with nominal – (13)a and (13)c and pronominal – (13)b and (13)d possessors (Vuković et al., forthcoming).

(13) a) *pomréše* *sví* *ostáde* *mi* *sámo* ***na*** ***bráta***
die.3PL.AOR all stay.3SG.AOR I.DAT.CL only on brother.OBL.SG
"All have perished, only I with my brother are left."

b) ***na*** ***méne*** *tétka* *umréla*
on I.OBL aunt.F.NOM.SG die.PPART.F.SG
"My aunt died."

c) *ozgór* *de* ***bojánu*** *kúťa* *bilá*
up there where bojan.M.DAT.SG house.F.NOM.SG be.PPART.F.SG
"You give to him, he (gives) to you."

d) ***méni*** je *májka* iz marinóvac
I.DAT AUX.3SG.PRES mother.F.NOM.SG from Marinovac.M.OBL.SG
"My mother is from Marinovac."

The existence of a clear standard equivalent to the dialectal features presented here allows us to measure the diastratic variation observed in the speech of single speakers and hence assess the degree to which speakers use the dialect. The operationalization of these features is illustrated in the next section.

3.7 Operationalization

Our analysis made use of frequencies for each feature described in Sections 3.2–3.6 to achieve two objectives:

i. Investigate whether there was a positive statistical correlation between any one feature and the other features, and ascertain which feature is the best predictor of the others. The initial hypothesis is that accent will be the best predictor.
ii. Segment and cluster the speaker-based occurrence of values in order to obtain groups ranked according to the distribution of dialectal vs. standard structures.

The analysis was performed on the variety of Torlak spoken in the Timok region in South-Eastern Serbia. Section 4 provides a detailed description of this sample.

4. The Timok sample

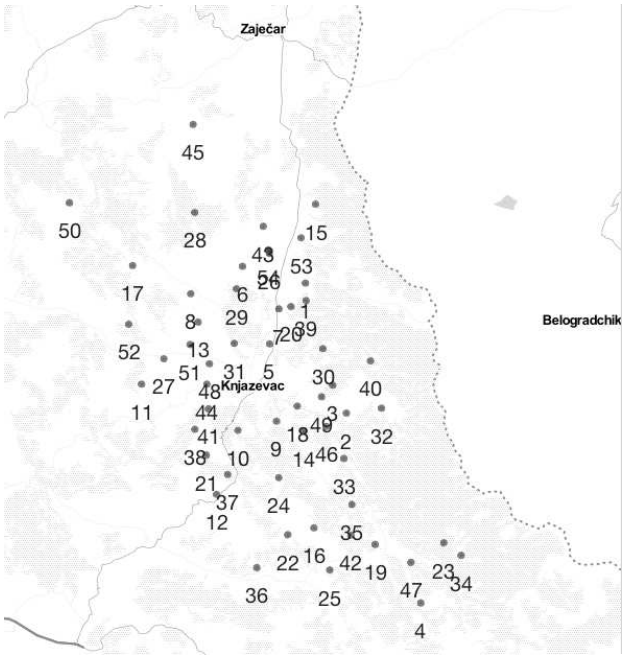
The sample used in the study is taken from the Spoken Torlak dialect corpus (Vuković, 2020),[6] which consists of fieldwork interviews conducted between 2015 and 2018 in the Timok region within Serbia's Prizren-Torlak dialect zone (Figure 1) (Vuković et al., forthcoming; Miličević-Petrović et al., forthcoming). Semi-structured interviews were conducted using an ethno-linguistic questionnaire (Sikimić, 2012; see Plotnikova, 1996) in order to elicit longer narratives on topics related to the culture and history of the region as well as biographical stories.

Figure 1. The Timok area within the Torlak dialect zone

[6] Available online at: http://hdl.handle.net/11356/1281.

In order to enable the geographical mapping of dialect features and the degree to which different speakers use Timok dialect, evenly distributed data points are included in the corpus. For each data point, at least one speaker was selected whose language was characterized as representative for the non-standard variety. The speakers were selected on the basis of the linguistic criteria that characterize Timok and the judgment of experts and native speakers with an understanding of the dialect. The entire corpus includes a total of 163 speakers (out of which some provided an insufficient amount of data for analysis, i.e. fewer than 1,000 tokens), as well as 12 researchers.

The present study includes only speakers older than 55 who contributed 1,000 tokens or more, and whose production provides data for the features under scrutiny. The resulting sample contains texts from 67 speakers representing 54 locations spread across the Timok area (Figure 2). This sample contains 385,517 tokens.



1	Ošljane	28	Marinovac
2	Drvnik	29	Debelica
3	Balinac	30	Šarbanovac
4	Ćuštica	31	Štipina
5	Gornje Zuniče	32	Aldinac
6	Trnovac	33	Pričevac
7	Jelašnica	34	Crni Vrh
8	Lepena	35	Gabrovnica
9	Žukovac	36	Guševac
10	Žlne	37	Krenta
11	Vasilj	38	Orešac
12	Tijovac	39	Petruša
13	Drečinovac	40	Radičevac
14	Gornja Sokolovica	41	Rgošte
15	Mali Izvor	42	Stara Kalna
16	Šesti Gabar	43	Vrbica
17	Bučje	44	Glogovac
18	Kandalica	45	Leskovac
19	Inovo	46	Gradište
20	Jakovac	47	Balta Berilovac
21	Ponor	48	Lokva
22	Vlahovo	49	Staro Korito
23	Janja	50	Borovac
24	Donja Kamenica	51	Bulinovac
25	Jalovik Izvor	52	Zorunovac
26	Minićevo	53	Selačka
27	Balanovac	54	Drenovac

Figure 2: Map of locations in Timok

Even though the majority of the speakers selected were born when primary education in standard Serbian was already obligatory and were exposed to the standard variety through mass media (mainly television), their speech is not uniform and displays gradual interspeaker variation. While such variation can be perceived for each individual sample, it cannot be measured precisely and consistently for the sample overall.

The transcription of interviews was semi-orthographic in order to preserve particular non-standard morpho-phonetic characteristics, features of the spoken language such as elisions, as well as non-verbal vocal content such as coughing or laughter. They also mark the position of the accent. Based on prosody, meaning, and structure, the transcripts for each speaker were split into meaningful segments, normally 1–6 seconds in duration. Pauses and hesitations, overlaps, and interruptions were also marked. An example of a transcribed text is shown in (14). In the example, utterance boundaries are marked with vertical bars. In transcripts, middle-dots are used to mark pauses; slashes are used for interruptions. For the purpose of the analysis, a plain text format was used. The corpus contains morphosyntactic annotation, following the MultextEast scheme (Erjavec, 2012), and lemmatization, which were not used in the analysis. The PoS tags and lemmas were assigned automatically using a customized model of the ReLDI tagger (Vuković, 2020; for original tagger see Ljubešić et al., 2016).

(14) *čovék • baštá ni je bío • prvi u • saló • gázda je bío || imáo imóvinu • • • i své imáo || níje kupuváo níšta || jedíno je nosío kom/ krómpir se mnógo rađáo || támo po ováj séla pírotska*
"The man, our father was the first one in the village || he was a landlord, he had property, he had everything || he did not buy anything || he was only carrying pot/ potatoes were growing in abundance || over there in these villages around Pirot" (Female, 75 years, Ravno Bučje, 2016)

5. Measuring variation

The extent of non-standardness in the selected sample was determined by measuring the frequency of the five selected features for each speaker. The obtained frequencies were then correlated with each other to identify the feature which could serve as the predictor of variation and as a baseline for non-standardness. The values for each speaker were used to calculate distances and similarities so as to distinguish the sub-samples that were, respectively, more dialectal and those that were more standard.[7]

[7] The files containing the Python and R scripts used in the analysis, as well as the data can be found in the GitHub repository at: https://github.com/bravethea/Timok_features_variation.

5.1 Analysis

Absolute frequencies were extracted using custom-made Python scripts that search for word forms based solely on the text. In the analysis, linguistic variables were expressed using frequencies that normalize the inconsistent lengths of transcripts.

As outlined above (see Section 3.2), non-standard accent counts were used as a non-standard feature. At the same time, they were tested in parallel with the other two features as possible general measures of non-standardness. Occurrences of non-standard accents were counted using a word list made up of entries with stress accents expected to vary from the standard variety. The full list has 94 words belonging to different categories (see Appendix 2 for the full list and Table 1 for some frequent examples). The number of words with non-standard accents was standardized for each text as a percentage against the total number of occurrences of words from the list.

When it comes to *si* as a dative clitic pronoun, all *si* word forms were extracted, including instances of *si* as a valency marker. The form si may also instantiate the second person present form of the verb *jesam/biti* ("to be"). These occurrences were manually removed from results. The values were normalized relative to the number of occurrences of *si* per 1,000 verbs. Note that even though frequency is commonly normalized per 1,000 or 10,000 words, this number has been decreased to 100 in order to scale the values, which would otherwise have been too low in relation to the other two variables.

In order to investigate the omission of the 3rd person auxiliary with perfect constructions, the search investigated at all past participle forms referring to the third person singular or plural[8] which did not have the third person auxiliary, such as *je, e* or *su*, in the three-word context before or one word after (based on 200 randomly picked examples, it was found that the context of this size covers most cases occurring in the corpus). The values are presented in percentages relative to 1,000 occurrences of the perfect tense.

Post-positive articles can be identified by their ending, therefore we searched for words with the typical demonstrative-like ending, such as *-at*, *-ta*, *-to*, and manually filtered the

[8] The number of occurrences of all but third person auxiliary forms in the left and right context was subtracted from the total number of occurrences of the past participle. This resulted in the total number of occurrences of the third person perfect. From this, the number of occurrences of the third person auxiliary forms was subtracted to obtain the final counts of the occurrences of bare participles.

unwanted results. The occurrences were then sorted according to speakers and normalized per 1,000 occurrences of nouns.

Searching for analytic case marking was the most challenging task. The data was initially extracted by searching for synthetic and analytic morphological forms or constructions, and then filtered manually (the manual data extraction and was performed by Mirjana Mirić for the purpose of the analysis in Vuković et al., forthcoming). In case of nouns, the forms of the synthetic dative were searched based on the inflectional endings for nouns (e.g. *ženi* / *ženama* "to the woman / women", *učeniku* / *učenicima* "to the pupil(s)". The analytic constructions were queried using a pattern *na* + *case*, where the case form was accessed using typical endings, allowing modifiers or non-verbal elements (pauses, laughter, or other interjections). Concerning pronouns, the search focused on the set of forms in dative or *na* + *oblique case*. Searches returned a large number of unwanted results given that these endings can also be found in many other forms and that similar constructions can perform other functions, since the preposition *na* is also used to mark location. Out of all the examples, only those with the IO and POSS were kept for analysis. Their frequency was normalized per 1,000 nouns for each speaker.

The descriptive statistics for each of the features is shown in Table 3. None of the variables are normally distributed. A visualization of their distribution across speakers is displayed in Figure 3.

Table 3: Quantitative values of the dialect features.

	Min	Median	Mean	Max	Standard deviation
Non-standard accent (proportion)	0%	65.52	62.42	92%	19.12
Pronoun *si* (norm. freq.)	0	25.33	30.54	88.13	19.13
Auxiliary omission (norm. freq.)	354.430	611.80	610.97	816.83	86.68
Post-positive article (norm. freq.)	0	7.44	16.16	84.28	20.18
Analytic case marking (proportion)	0%	100	95.95	100%	13.65

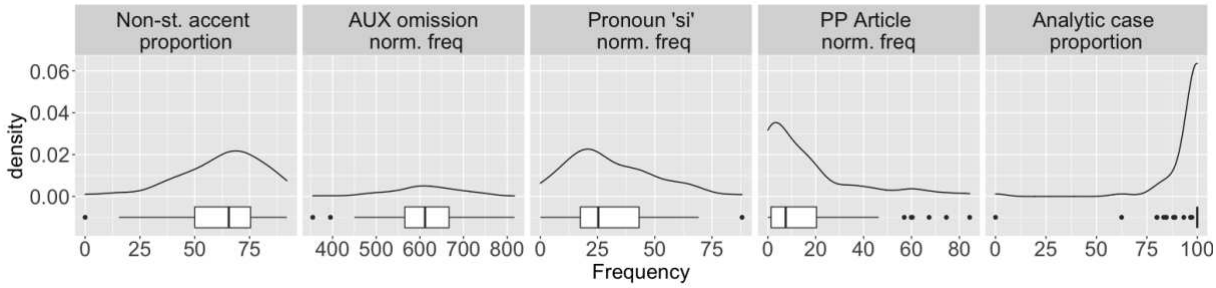


Figure 3: The distribution of the dialectal features across speakers (N = 67)

As previously stated, the chosen features are not expected to reflect – from a purely synchronic point of view – the areal influence of neighboring languages, but rather demonstrate vertical contact with Serbian. In order to verify this claim, the frequencies were correlated with longitude and latitude, which are chosen as proxies for the contact with East/West or North/South and correlated with the linguistic features using Pearson's correlation. The results do not show a relationship between AUX omission, the use of the pronoun *si* and the analytic case marking, or the geographic coordinates, so a horizontal influence of Serbian or Bulgarian[9] on the use of these features does not seem likely. A positive effect of longitude exists on the post-positive article ($r=0.50$, $df=98$, $p<0.001$), and a marginal negative effect of the latitude ($r=-0.28$, $df=98$, $p<0.05$). There is a marginal negative effect of longitude on the non-standard accent ($r=-0.23$, $df=98$, $p< 0.05$), but not of the latitude. The geographical pattern can easily be observed in the maps given in Figure 6, where the frequency is represented by the size of the circle. Because geographic variables cannot be used as a consistent indicator of the variation, also because sometimes speakers from the same location display different properties, we are approaching the variation based on features alone.

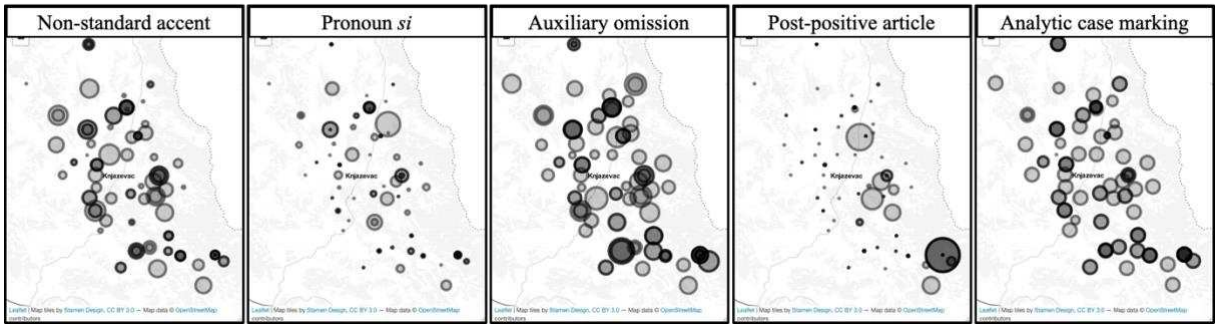


Figure 4: Maps of the geographical distribution of the features

[9] Here it was assumed that language boundaries are identical to the political boundaries between the two countries. Given the political situation throughout most of the 20th century, this seems a fair assumption.

In order to investigate the relationship between the features, the occurrences of each feature were correlated with the other features in pairs, using Pearson's correlation in R, whereby we hypothesize that the accent will surface as the best predictor of the other features. This tests if there is a positive correlation and hence whether accent is indeed less voluntarily controlled and hence verifies our initial hypothesis that it can predict a speaker's overall production relative to the influence of the standard variety.

The second course of our analysis was set to identify levels of non-standardness in the Timok sample. In order to identify the speakers with the most characteristically dialectical speech, those with the greatest frequency of dialect features were selected. For this, an unsupervised method of hierarchical clustering was applied that groups individual samples automatically based on distance. The group with the highest mean values was then taken as the most non-standard and vice versa. Three clusters was chosen based on a visual assessment of the hierarchical dendrogram and by inspecting the values obtained for each cluster (see Section 5.2). The distance matrix was derived using a Euclidean distance measure. Clustering was performed using Ward's minimum variance method.

5.2 Results

The results of a Pearson's correlation test show that none of the five features correlates with all four of the others. Table 4 shows the scores of the correlation tests between the features. The non-standard accent and the auxiliary omission correlate with the three other features. The non-standard accent correlates positively with the pronoun *si*, the AUX omission and the analytic case marking, but not the post-positive article. The auxiliary omission correlates positively with the accent, the post-positive article, and case. The other features correlate only with two features each (see Table 4). The strongest correlation is between the pronoun *si* and the post-positive article.

Table 4: Correlation levels of Pearson's correlation between the features (N=67, df=65)

	accent	si	AUX	article	case
accent		r=0.25 p<0.05	r=0.29 p<0.02	r=0.16 p=0.21	r=0.25 p<0.04
si	r=0.25 p<0.05		r=0.22 p<0.07	r=0.39 p<0.001	r=0.10 p<0.5

AUX	r=0.29 p<0.02	r=0.22 p<0.07		r=0.26 p<0.04	r=0.27 p<0.03
article	r=0.16 p=0.21	r=0.39 p<0.001	r=0.26 p<0.04		r=0.15 p<0.21
case	r=0.25 p<0.04	r=0.10 p<0.5	r=0.27 p<0.03	r=0.15 p<0.21	

The hierarchical clustering method, visualized in the dendrogram in Figure 7, divides the data into 3 groups, each with a mean value, as presented in Table 5. Speaker codes marked at the end of the branches can be used for sub-sampling the corpus data (see Appendix 3). For purposes of presentation, the frequency values were scaled between 0 and 1 in order to decrease the large differences between the values in each variable (analysis was performed on unscaled values). Note that cluster labels are randomly assigned, and their numerical order does not refer to the level of non-standardness. The mean values visualized in Figure 8 indicate that cluster 1 is the most dialectal, while cluster 2 is the least dialectal and cluster 3 stands in the middle between the other two.

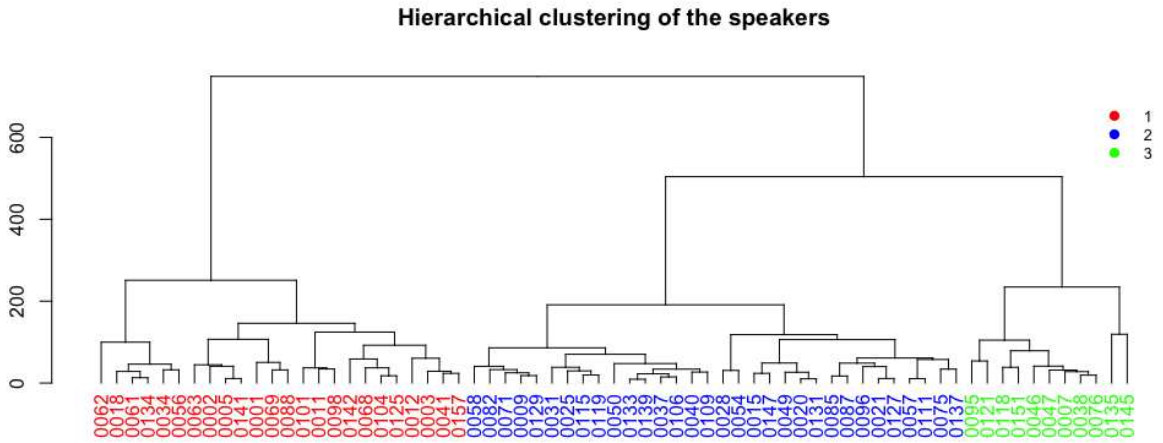


Figure 5: Dendrogram representing the hierarchical clustering of the speakers using all five dialectal features.

Table 5: Clusters size and mean values

Cluster	**N**	**Non-standard accent**	**Pronoun *si***	**AUX omission**	**Post-positive article**	**Analytic case marking**
Cluster 1	24	0.74	0.31	0.71	0.98	0.43
Cluster 2	11	0.25	0.09	0.54	0.89	0.2
Cluster 3	32	0.52	0.14	0.7	0.97	0.33

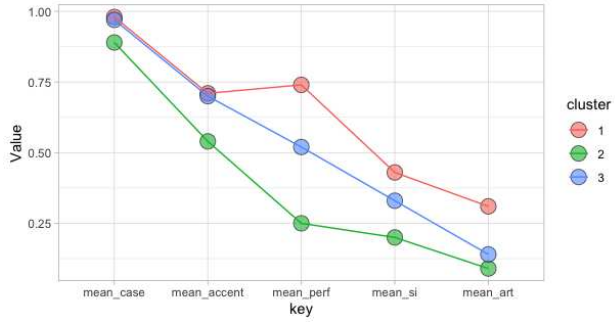


Figure 6: Mean cluster values for each feature.

The map showing the geographic distribution of the clusters (Figure 10) does not reveal a clear geographic distribution pattern.

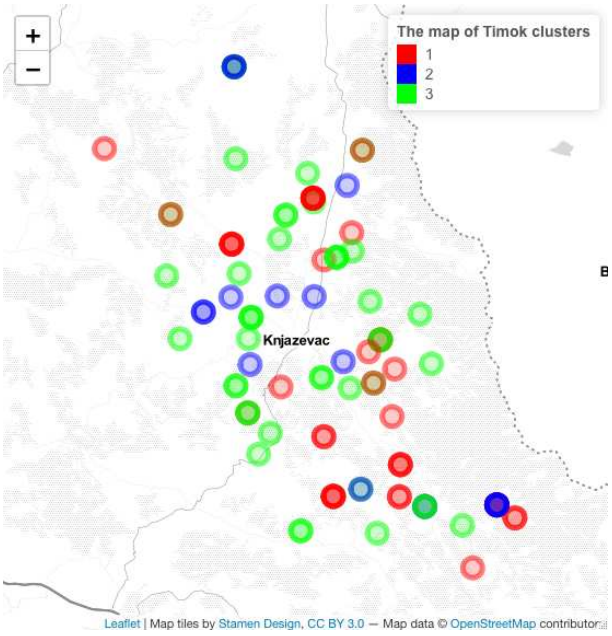


Figure 7: Map with Timok clusters (n = 41)

6. Discussion

Results indicate that, in the sample used for the analysis, the use of the non-standard accent and the omission of the auxiliary in the perfect tense are the best predictors of non-standardness for dialect texts from Timok among the features taken into consideration. However, this inference is made on the basis of the analysis of merely five features, and should be further confirmed with other dialect features and tested against intuitive selection.

The model validates our assumption that accent would reveal dialectal speakers because it is more difficult to self-correct than other salient features. Auxiliary omission in the perfect tense is a feature that is not so striking given that it is, to a lesser extent, present in colloquial Serbian. As a result, it is less prone to self-correction.

The fact that the pronoun *si*, the post-positive article, and the analytic case marking correlate to a lesser extent with the other features points to their salience and the self-awareness of the speakers with respect to their use. This again confirms that very salient features cannot be used alone as general indicators of variation. This kind of insight is also useful when it comes to overcoming the observer's paradox, which often arises when the researcher is not fluent in the dialect and which can have an influence on the use of salient dialectal features.

The approach to measuring variation proposed here relies on linguistic features identified as markers of the dialect as opposed to the standard language. As such, the method proposed can be used for a corpus-based study of dialectal variation, representativeness and language vitality beyond the specific case of Torlak, using an adapted set of relevant features.

In order to gain insight into the triggers underlying the variation observed, factors such as socio-demographic parameters, e.g. age, gender, education, or geographic factors, such as longitude, latitude, altitude, or mixed factors such as isolation or distance from urban centers, need to be considered.[10] A contrastive analysis based on the data provided in Belić (1905), Stanojević (1911) and Sobolev (1998) and, if possible, even older data, will provide insight into whether the situation evidenced by the data indeed reflects a change as compared to older times. This promises to provide more insights into feature diffusion, as well as with respect to horizontal contacts with neighboring Bulgarian and Macedonian dialects.

7. Conclusion

This paper approached the analysis of variation in a corpus of the Torlak dialect, a dialect which is, in varying degrees, influenced by standard Serbian. It analyzed the frequency of occurrence of five dialect features of the Timok dialect of Torlak in a sample of 67 speakers: non-standard accent position, the clitic pronoun *si,* omission of the 3rd person auxiliary in the perfect tense, the post-positive article, and the analytical dative marking of the indirect object and possessive.

[10] This was shown to be significant in the analysis of the use of post-positive articles by Vukovic and Samardžić (2018).

Taking the Torlak dialect and standard Serbian as two poles of a variation scale, it aimed at empirically distinguishing between gradual levels of non-standardness and to establish accent position as a predictor for the use of other dialect features.

Using the method of hierarchical clustering, the most standard-like and the most Torlak-like varieties were identified. Using Pearson's correlation method, it was demonstrated that accent and auxiliary omission can serve as an indication of non-standardness in the Timok sample. At the same time, the inability to identify a single reliable predictor of non-standardness illustrates the complexity of Timok's linguistic variability.

The results of the methods presented in this paper will be used for classification of the corpus data, as well as to identify data on different poles that may be used for contrastive analysis of the use of dialect. Such analysis can be expanded to include more features of Torlak to further test the hypotheses.

Furthermore, this method can be used for the stratification of speakers in situations of vertical feature transfer (from standard language to a dialect), where the set of features affected, and their manifestations, are formally identifiable. In a similar manner, diachronic developments could be analyzed where the change or grammaticalization of features could be witnessed in speech over different generations. Feature-based clustering of speakers from different historical periods could be performed empirically and compared to common theoretical assumptions. As such, the work presented here paves the way to creating a method for incorporating data-driven empirical information on variation in dialect corpora.

Acknowledgments

The study was supported by the *project (Dis-)entangling traditions in the Central Balkans: Performance and perception (TraCeBa)* funded by ERA.Net RUS Plus and the Swiss National Science Foundation, and the project *'Ill-Bred Sons,' Family and Friends: Tracing the Multiple Affiliations of Balkan Slavic*, funded by the Swiss National Science Foundation (SNF 100015_176378/1).

References

Alexander, R. (1975). *Torlak Accentuation.* Otto Sagner.

Arsenijević, B. (2012). Evaluative Reflexions: Evaluative Dative Reflexive in Southeast Serbo-Croatian. In B. Fernandez, & R. Etxepare (Eds.), *Variation in Datives: A Microcomparative Perspective* (pp. 1–21). Oxford University Press.

Belić, A. (1905). *Dijalekti istočne i južne Srbije*. Srpska Kraljevska Akademija.

Bruland, I., & Carr, P. (2013). Variability, unconscious accent adaptation and sense of identity: The case of RP influences on speakers of Standard Scottish English. *Language Sciences*, *39*, 151–155.

Chambers, J. K., & Trudgill, P. (1998). *Dialectology*. Cambridge University Press.

Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources & Evaluation*, *46*, 131–142. https://doi.org/10.1007/s10579-011-9174-8

Escher, A. (2021). Double argument marking in Timok dialect texts (in Balkan Slavic context). *Zeitschrift für Slawistik*, *66*(1), 61–90.

Escher, A. (2021). Auxiliary Omission in the Perfect Tense in Timok. *Balkanistica*, *34*, 41–64.

Frleta, T. (2010). Uporaba i značenje nenaglašenog dativa povratne zamjenice u hrvatskom jeziku. *Jezik: časopis za kulturu hrvatskoga književnog jezika*, *57*(1), 1–13.

Grickat, I. (1954). *O perfektu bez pomoćnog glagola u srpskohrvatskom jeziku*. Srpska Akademija Nauka.

Hinrichs, U. (1999). Die sogenannten Balkanismen als Problem der Südosteuropa Linguistik und der Allgemeinen Sprachwissenschaft. In U. Hinrichs (Ed.), *Handbuch der Süosteuropa-Linguistik* (pp. 42–463). Harrassowitz.

Ivanova, E. Y., & Gradinarova, A. A. (2015). *Sintaksicheskaya imok bolgarskogo yazyka na fone russkogo*. Yazyki slavyanskoj kultury.

Ivić, P. (1985) *Dijalektologija srpskohrvatskog jezika. Uvod i štokavsko narečje* [Dialectology of the Serbo-Croatian Language. Introduction and the Neo-Shtokavian dialects]. Matica srpska.

Ivić, P. (2009). *Srpski dijalekti i njihova klasifikacija*. Izdavachka knizharnitza Zorana Stojanovicha.

Joseph, B. (1992) The Balkan Languages. In W. Bright (Ed.), *International Encyclopedia of Linguistics* (Vol.1, pp.153–155). Oxford University Press.

Krstić, D. (2014). *Konstrukcija identiteta Torlaka u Srbiji i Bugarskoj* [Doctoral dissertation]. Univerzitet u Beogradu.

Lindstedt, J. (2000). Linguistic balkanization: Contact-induced change by mutual reinforcement. In: D. Gilbers, J. Nerbonne, & S. Schaeken (Eds.), *Languages in Contact: Studies in Slavic and General Linguistics* (pp. 231–246). Rodopi.

Ljubešić, N., Klubička, F., Agić, Ž., & Jazbec, I. (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference*

on Language Resources and Evaluation: LREC 2016 (pp. 4264–4270). European Language Resources Association.

Meerman, A., & Sonnenhauser, B. (2016). Das Perfekt im Serbischen zwischen Slavischer und Balkanslavischer Entwicklung. In A. Bazhutkina & S. Barbara (Eds.), *Linguistische Beiträge zur Slavistik. XXII. Jungslavistlnnen-Treffen in München. 12. Bis 14. September 2013* (pp. 83–110). Biblion Media.

Miličević Petrović, M., Vuković, T., Mirić, M., Konior, D., & Escher, A. (forthcoming). Language Documentation II: Towards a sociolinguistic corpus of Torlak. Challenges for data processing. *Zeitschrift für Slavische Philologie*. Winter.

Mišeska-Tomić, O. (2004). The Balkan Sprachbund Properties. In O. Mišeska Tomić (Ed.), *Balkan Syntax and Semantics* (pp. 1–55). John Benjamins.

Mitkovska, L. (2011). Competition between nominal possessive constructions and the possessive dative in Macedonian. In M. Nomachi (Ed.), *The Grammar of Possessivity in South Slavic languages and Diachronic Perspective* (pp. 83–109). Slavic Research Center.

Nerbonne, J., & Kretzschmar, W. A. (2012). Dialectometry ++. *LLC: Journal of Digital Scholarship in the Humanities*, *28*(1). pp. 2-12.

Petrova, G. (2014). Medialny glagoly s refleksivna semantika. *Nauchny trudove na Rusenskyja universitet. Serija*, *53*(6.3), 36–40.

Petrović, T. (2015). *Srbija i njen jug. "Južnjački dijalekti" između jezika, culture i politike*. Fabrika knjiga.

Plotnikova, A. A. (1996). *Materialy dlja etnolinggvisticheskogo izuchenija balkanoslavjanskogo areala*. Institut slavjanovedenija RAN.

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Savova, D. (2017). Glagoli s elementa si/sobie v balgarskija i v polskija ezik. *Zeszyty Cyrylo-Metodiańskie*, *6*, 38–56.

Schmidt, T. (2009). Creating and working with spoken language corpora in EXMARaLDA. In V. Lyding (Ed.), *Proceedings of the Second Colloquium on Lesser Used Languages & Computer Linguistics* (pp. 151–164). EURAC research.

Sikimić, B. (2012). Timski terenski rad Balkanološkog instituta SANU. Razvoj istraživačkih ciljeva i metoda. In. M. Ivanović-Barušić (Ed.), *Terenska istraživanja – poetika susreta* (pp. 167–198). Etnografski institut SANU.

Szmrecsanyi, B. (2015). *Grammatical variation in British English dialects: A Study in Corpus-Based Dialectometry*. Cambridge University Press.

Szmrecsanyi, B. (2017). Variationist sociolinguistics and corpus-based variationist linguistics: overlap and cross-pollination potential. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, *62*(4), 1–17.

Szmrecsanyi, B., & Wälchli, B. (Eds.) (2014). *Aggregating Dialectology, Typology and Register Analysis: Linguistic Variation in Text and Speech*. Walter de Gruyter.

Sobolev, A. N. (1998). *Sprachatlas Ostserbiens und Westbulgariens. Bd. III. Texte*. Biblion Verlag.

Sobolev, A. N. (2003). *Malyj dialektologicheskij atlas balkanskih jazykov. Probnyj vypusk*. Biblion Verlag.

Sobolev, A. (2008). From synthetic to analytic case: Variation in South-Slavic dialects. In A. Malchukov, & A. Spencer (Eds.), *The Oxford Handbook of Case*. Oxford University Press.

Stanojević, M. (1911). Severno-timočki dijalekat. *Srpski dijalektološki zbornik*, *2*, 360–463.

Stevanović, M. (1986). *Gramatika srpskog jezika*. Naučna kniga.

Trudgill, P. (1986). *Dialects in Contact*. Blackwell Publishers.

Vuković, T. (2019). *Torlak ReLDI Tagger 2019* [Computer software]. Retrieved November 1, 2021, from https://github.com/bravethea/Torlak-ReLDI-Tagger-2019.

Vuković, T. (2020). *Spoken Torlak dialect corpus 1.0*. CLARIN.SI. http://hdl.handle.net/11356/1281.

Vuković, T. (2021). Representing variation in a spoken corpus of an endangered dialect: The case of Torlak. *Language Resources & Evaluation*, *55*, 731–756.

Vuković, T., Mirić, M., Escher, A., Ćirković, S., Miličević Petrović, M., Sobolev, A., & Sonnenhauser, B. (forthcoming). Under the magnifying glass. Dimensions of variation in the contemporary Timok variety. processing. *Zeitschrift für Slavische Philologie*. Winter.

Vuković, T., & Samardžić, T. (2018). Prostorna raspodela frekvencije postpozitivnog člana u timočkom govoru. In S. Ćirković, A. N. Sobolev, B. Sonnenhauser, M. Miličević, & J. Pandurević, (Eds.), *Timok. Folkloristička i lingvistička terenska istraživanja 2015–2017* (pp. 181–200). Narodna biblioteka "Njegoš".

Wahlström, M. (2015). *The loss of case inflection in Bulgarian and Macedonian* [Doctoral dissertation, University of Helsinki]. University of Helsinki, Department of Modern Languages. https://researchportal.helsinki.fi/en/publications/loss-of-case-inflection-in-bulgarian-and-macedonian.

Appendix 1

As an additional analysis, we compared how younger speakers in the corpus make use of non-standard features with the older ones that are known to be more dialectal. The sample of younger speakers consists of 10 interviews with high-school students, all living in Knjaževac. Their language is closer to the standard because of the impact of standard Serbian taught in school (Vuković et al., 2020), which was also predicted by the native speakers.

The comparison of the two samples shows a notable difference in the use of the three features by the older and the younger population. Note that in the younger group, only three

speakers provided examples for the analytic case marking. In the younger population, the mean value for the non-standard accents is 19.19, for the use of the pronoun *si*, it is 7.40, for the omission of the AUX 556.82, for the post-positive article 3.81, and for the analytic case marking 33.33. These values are considerably lower than the respective values in the sample used for the primary analysis: non-standard accent 62.42, pronoun *si* 30.54, AUX omission 610.97, post-positive article 16.16, and analytic case marking 95.95. The compared pairs of values between the two samples are presented in boxplots in Figure 8.

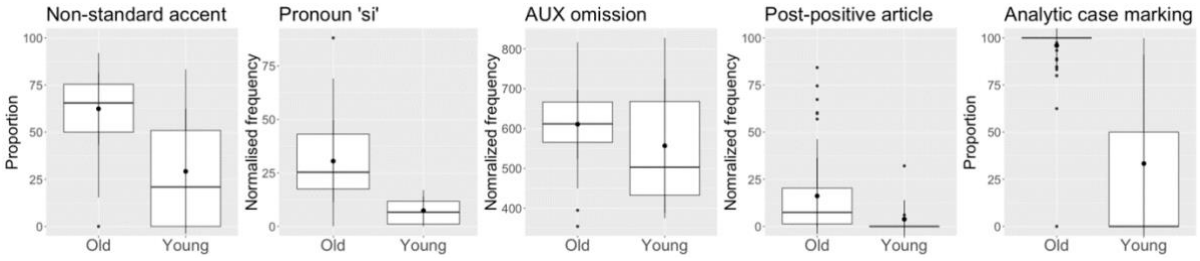


Figure 8: Comparison between older and younger speakers

The results of the evaluation demonstrate that the method yields the expected results from the data – an obvious distinction in the level of use of dialect features. However, this kind of differentiation cannot be achieved using the factor of age within the older group.

Appendix 2

The list of words used in for the extraction of non-standard accents:

brazda, brazdu, voda, vodu, glava, glavu, greda, gredu, Timok, zvezdu, zemlja, zemlju, zima, zimu, zora, zoru, žena, ženu, igla, iglu, koza, kozu, kosa, kosu, magla, maglu, metla, metlu, noga, nogu, ovca, ovcu, reka, reku, ruka, ruku, svinja, svinju, sveća, sveću, sveča, sveću, sestra, sestru, snaja, snaju, torba, torbu, međa, medža, među, medžu, planina, planinu, deca, decu, dete, mleko, čovek, ručak, čovek, čovek, čovek, krstovi, kakvo, tako, bila, bilo, bili, bile, jedan, edan, jedwn, jedna, jednu, u,nesi, nesu, nisi, nisu, nesam, neswm, nisam, niswm, išla, išal, išli, išlo, unuk, kako, koji, moja, tvoja, tva

Appendix 3

The list of speaker codes segmented in clusters. Speaker labels in the corpus start with TIM_SPK_ followed by the code.

Cluster 1:
0001 (Ošljane), 0002 (Drvnik), 0003 (Balinac), 0005 (Ćuštica), 0011 (Jelašnica), 0012 (Lepena), 0018 (Žlne), 0034 (Mali Izvor), 0041 (Bučje), 0056 (Ponor), 0061, 0062 (Vlahovo), 0063 (Janja), 0068, 0069 (Donja Kamenica), 0088 (Lepena), 0098 (Pričevac), 0101 (Crni Vrh), 0104 (Gabrovnica), 0125 (Stara Kalna), 0134 (Gradište), 0141 (Staro Korito), 0142 (Borovac), 0157 (Drenovac);

Cluster 2:
0007 (Gornje Zuniče), 0038 (Šesti Gabar), 0046 (Kandalica), 0047 (Inovo), 0076 (Balanovac), 0095 (Štipina), 0118 (Janja), 0121 (Rgošte), 0135 (Leskovac), 0145 (Bulinovac), 0151 (Selačka);

Cluster 3:
0009 (Trnovac), 0015 (Žukovac), 0020 (Vasilj), 0021 (Tijovac), 0025 (Drečinovac), 0028 (Gornja Sokolovica), 0031 (Mali Izvor), 0037 (Šesti Gabar), 0040 (Bučje), 0049 (Inovo), 0050 (Jakovac), 0054 (Balinac), 0057, 0058 (Ponor), 0071 (Jalovik Izvor), 0075 (Minićevo), 0082 (Marinovac), 0085 (Debelica), 0087 (Šarbanovac), 0096 (Aldinac), 0106 (Guševac), 0109 (Krenta), 0111 (Orešac), 0115 (Petruša), 0119 (Radičevac), 0127 (Vrbica), 0129 (Glogovac), 0131 (Leskovac), 0133 (Gradište), 0137 (Balta Berilovac), 0139 (Lokva), 0147 (Zorunovac).