Universität
Zürich UZH

Empirical Approaches to Variation. The Case of Timok Variety of Torlak

Thesis (cumulative thesis)
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of Doctor of Philosophy

by
Teodora Vuković

Accepted in the spring semester 2022
on the recommendation of the doctoral committee composed of

Prof. Dr. Barbara Sonnenhauser (main supervisor)

Prof. Dr. Hanne Eckhoff

Zurich, 2024

Acknowledgements

First and foremost, I want to express my greatest gratitude to my doctoral supervisor, Barbara Sonnenhauser. She was a better supervisor than I could have imagined: incredibly dedicated and thoughtful. She motivated and supported me, and wielded a great influence throughout the long years of my doctorate that have left a profound mark on me. Given the extensive advice she provided, she should be considered a rightful co-author of all the articles in this dissertation.

I extend my gratitude to my second supervisor, Hanne Eckhoff, who set a great example with her own work that I extensively followed. She provided excellent technical feedback, following her diverse linguistic expertise, and offered valuable career advice when needed. I owe special thanks to Tanja Samardžić for her support in helping me establish my first connections in Zurich, for helping me gain solid foundations in programming and computational linguistics, and for teaching me important life lessons.

I am extremely thankful to my dear colleagues from Serbia, who took me on board the journey of Balkan linguistics. Biljana Sikimić invited me to join the team on the journey in Timok and helped me write my first project proposal, which ultimately became this dissertation. I owe gratitude to Svetlana Ćirković and Mirjana Mirić for companionship, advice, wonderful years of collaboration, and support with data processing, which resulted in the Torlak corpus. My appreciation goes to Maja Miličević Petrović for mentoring me toward my MA degree and sharing her expertise during the TraCeBa project. Our Serbian fieldwork expeditions were accompanied by our dear colleague, Max Wahlström, who introduced me to many relevant linguistic concepts, approaches to research, and inspiring colleagues. I thank Kristina Eiviler for our amazing days in Timok, her efforts in transcription, and for being an inspiring colleague. Most of all, I thank her for the many laughs and for being a good friend to this day.

My work benefited greatly from collaboration with our Russian colleagues, Andrey Sobolev, Daria Konior, and Anastasia Escher. More than just a project partner and a collaborator in some of the most satisfying research, I am deeply thankful to Nastia for being a close friend, and for many climbing sessions and long conversations. Further project contributions were made by Sanja Brađan, Danijela Stojković, and Bojana Damnjanović, who transcribed and annotated the dialect texts.

Gratitude is also due to Ivan Šimko and Olivier Winistörfer at the Slavisches Seminar, who created a wonderful working atmosphere and actively participated in numerous research exchanges. Additionally, Florian Wandl not only made excellent coffee but also participated in my doctoral defense. I was fortunate to have worked surrounded by such great people.

My most sincere gratitude goes to all the reviewers of my articles for their invaluable feedback, as well as to the editorial teams of the following journals that contributed their efforts to the papers in this dissertation: Language Resources and Evaluation, International Journal of Corpus Linguistics, Zeitschrift für Slavische Philologie, and the Journal of Slavic Linguistics.

I thank my mother, Gorana, for supporting my academic ambitions, and Tjade for making me tea and supporting me through the final days of my doctorate.

The research in this dissertation was made possible by funding from the SNSF (grant agreement IZRPZ0_177557/1), the Russian Foundation for Basic Research (RFBR No. 18-512-76002), and the Serbian Ministry of Education, Science and Technological Development (grant agreement No. 401-00-00642/2018-09) with the Horizon 2020 ERA.Net RUS Plus program. This funding supported my employment at the Slavisches Seminar at the University of Zurich from 2018 to 2021, enabling collaboration within the TraCeBa project. Additionally, the data collection in Serbia within the project "The Protection of the Immaterial Heritage of the Timok Vernacular" between 2015 and 2017 was partially supported by a grant from the Ministry for Culture and Information of the Republic of Serbia. I was also honored to receive the Swiss Government Excellence Scholarship from 2016 to 2019, which facilitated my academic development and introduced me to many inspiring scholars worldwide, some of whom became close friends. Last but certainly not least, my work at the University of Zurich has been supported by the URPP Language and Space, where I worked and enjoyed many fruitful interactions, collaborations, and friendships. I am deeply grateful for all this support.

1. Introduction

The present dissertation delves into the intricate world of language variation, focusing specifically on the Timok dialect, a sub-standard variety spoken in Southeast Serbia and part of the larger Torlak dialect cluster. The Torlak dialects occupy a unique position at the crossroads of the West South Slavic and Balkan Slavic dialects, influenced by a range of factors including geographical, political, social, and historical circumstances. These influences are crucial for understanding the multifaceted nature of the dialect and its variations[1].

When I began this research with fieldwork in Timok in 2015, the landscape of linguistic research in this area was markedly different from what it is today. The field was characterized by a notable scarcity of contemporary empirical studies and electronic data on this linguistic area. Knowledge of the Torlak dialect, including its Timok variant, relied primarily on traditional dialectological descriptions, leaving a significant gap in our understanding of its sociolinguistic and areal complexities. This dissertation and surrounding research team efforts represent significant progress in research on the Torlak dialect, attributed to the joint efforts of an international team of linguists from Switzerland, Serbia, Russia, and Finland. Their collaborative work has been crucial in shedding new light on this linguistically rich yet under-researched dialect. During this time, research on the Torlak dialect has been supported by several funding sources (see Acknowledgements), which enabled extensive fieldwork in the Timok region and facilitated the creation and analysis of electronic data and empirical studies that were previously lacking, including the work presented here.

This cumulative dissertation consists of four distinct yet interconnected articles contributing to this endeavor, introducing new resources that have been developed and covering analyses tackling various facets of linguistic variation in Timok. Linguistic analyses focus on several linguistic features in which variation is present and use them to analyze areal and social variation. In terms of methodology, this study combines a quantitative variationist and corpus linguistic approach, using a corpus of authentic

[1] A comprehensive description of the current Torlak situation is given in Sobolev et al (2023), Sikimić et al (2023), that both appeared in the thematic issue of the *Zeitschrift für Slavische Philologie* (Volume 79, Issue 1) about Torlak. Many other publications about Torlak have been produced by the team to this date: Sobolev et al (2023), Sikimić et al (2023), Miličević et al. (2023), Konior et al. (2023), Ćirković et al. (2023), Mirić et al. (2021), Mirić (2017, 2018a, 2018b), Ćirković (2021, 2018a, 2018b, 2017), Escher & Vukovic (2020) (likely an incomplete list).

spoken data combined with relevant metadata to unravel the interplay of factors contributing to language variation in Timok.

In the following sections, a general background on the phenomenon of language variation is offered, with a particular emphasis on the current linguistic contact scenario within the Timok variety. Furthermore, the theoretical foundations and methodological approaches underpinning this study are presented, as well as the core concepts and insights of each study mapping the trajectory of this research. Finally, this text highlights the significant findings of these analyses, emphasizing their broader relevance and the contribution they make to our understanding of linguistic variation and change in Timok.

2. Phenomenon of Language Variation and Underlying Factors

Language variation is a constant, dynamic process in living languages. The small differences observed in individual varieties are one of the crucial components of language and its potential to adapt (D'Arcy & Tagliamonte, 2015: 1; Chambers & Trudgill, 2003, 1998; Wardhaugh & Fuller, 2014; Nerbonne, 2009; Tagliamonte, 2011, 2006: 6; Labov, 1972). Dynamics of variation across larger areas, time spans, or social groups yield notable differences in language, resulting in sociolects, dialectal or language continua, and language change.

Within a linguistic landscape, boundaries, be they natural, political, mental, or social, delineate communities of speakers and create environments where unique linguistic varieties develop. On one hand, boundaries often limit interaction and exposure to external linguistic influences, supporting internal variation within varieties and differentiating them from others. On the other hand, unlike administrative boundaries, which are clearly defined, linguistic boundaries are vague, characterized by gradual and sometimes subtle distinctions between varieties. Language boundaries are, in practice, often based on political and administrative borders rather than on linguistic differences or mutual unintelligibility. The isoglosses of linguistic features do not necessarily coincide with state borders. Nonetheless, it is important to "recognize that political or linguistic borders are not absolute barriers but may act as bridges that may be crossed" (Palander, Riionheimo & Koivisto, 2018: 7-14).

While written language, especially the standard variety, is “highly uniform and governed by prescription, the vernacular is most revealing of structured heterogeneity” (Weinreich, Labov & Herzog, 1968, as cited in D’Arcy & Tagliamonte, 2015: 1). Mechanisms of variation function more directly and obviously in spoken language, which is less affected by normative rules (Labov, 1984). Speakers always have the possibility to choose between different formal means of expression, while in spontaneous speech, they have less time to consider that choice (e.g., Kroch, 1978: 28). The most salient non-standard features would typically not appear in writing because speakers are aware of their stigmatization (Schneider, 2003: 59; Kautzsch, 2000: 222). Because of such circumstances, spoken non-standard vernaculars display the greatest amount of variation and can be considered a more natural manifestation of language (Kroch, 1978: 27-28; see Anderwald, 2011).

However, this variability is not unbounded; it is rather governed by a fine set of constraints. These constraints can sometimes be specific to a particular language, language situation, grammatical feature, or speakers’ environment. The goal of variationist linguistics is to analyze and decompose linguistic features as bundles of information into their building components on the formal surface and sub-surface level and establish tendencies and rules based on the internal structure of language, as well as to find which external factors may determine the choice of one linguistic pattern over the other.

Often, variation can be explained through internal structural aspects of language. An example of variation driven by language-internal mechanisms is the reduction in word-final consonant clusters variably simplified by deleting a coronal stop in certain English varieties, e.g., cost me > cos’ me (as in Anttila, 2002, based on studies by Labov et al., 1968; Wolfram, 1969; Fasold, 1972). This phenomenon arises as a way of facilitating pronunciation, embodying the principle of economicity or simplifying speech (Leopold, 1930; cf. Zipf, 1949; Tauli, 1958). On the other hand, there are limitations to this simplification phenomenon, which come into action when the simplification produces homonymy, making the differentiation between frequently used elements of language difficult and thus resulting in impaired comprehension. In that case, the principle of distinctiveness (ibid.) becomes important for maintaining the necessary level of understanding. In the previous example, there are two limiting factors. The deletion rate depends on the phonetic quality of the following segment; e.g., the reduction is more frequent before a fricative and less frequent before vowels

(Labov, 1997). Another factor is the morphological status of the segment: the deletion rate is more frequent with monomorphemes, such as cost, than when the cluster is a part of the regular past marker, as in tossed, because the morpheme is important for marking the distinction between marked past form and unmarked present tense form (Guy, 1994; Guy, 1991; Santa Ana, 1992).

Just as language-internal factors are important, external ones, such as the geographic and social environment, can play a significant role in how these patterns of variation, once started, disperse across space and speakers. “The social and geographical intersections of people’s daily life paths create contexts where accommodatory linguistic behavior can occur, but [there are] contexts where, for whatever reason, contact is – or has been – limited, resulting in dialect boundaries” (Britain, 2010: 209). The proximity and social status of surrounding varieties or languages, the shape of geographic terrain, or geopolitical infrastructure influence interaction between speakers and consequently grammars. The basic assumption in areal linguistics and dialectology is that geographically close varieties are more similar, and vice versa (Hickey, 2013; Matras, 2009; cf. Jeszenszky et al., 2017). Within this general rule, although gradual and increasing, the distancing between varieties is not uniform and directly correlated to physical distance alone. The diversity, complexity, and the diffusion of grammatical features depend on the population density, which is connected to other natural conditions, such as relief, water bodies, soil properties, and vegetation (Axelsen & Manrubia, 2014; Nichols, 2015), as well as by political borders and civilization infrastructure: traffic, settlement size, together with access to education, cultural institutions, and industry (Gooskens, 2005; Höder, 2011).

Typically, speakers conform to presupposed identities embodied in language; they make adjustments in their language, create, maintain, or decrease the social distance in interaction (Giles & Ogay, 2007), which in this case equates to linguistic distance. Thus, the social environment is another key influence on language variation. It is constructed through the social organization of society and collective ideas that communities share about sub-groups of that society and their mutual relationships. One of the dominant factors, especially in industrialized societies, is age. Younger people tend to adapt more and therefore use more progressive language features, whereas older people often tend to use more conservative ones (Chambers, 2003: Chapter 3; Chambers & Trudgill, 1998: 78–81). Another factor is gender, while the direction in which it affects language depends on the culture. In some situations, women have a more progressive language, while in Slavic and Eastern Europe, women are known to display more archaic features, resulting from the difference in education and migration compared to men (Chambers & Trudgill, 1998: 84-85; Petrović, 2015). In the context of social stratification, the concept of prestige plays an important role, which is often manifested when comparing standard and non-standard varieties (Chambers & Trudgill, 1998: 3, 69–76, 81–86; Ash, 2003). Standard varieties are often considered highly prestigious, while non-standard varieties are less prestigious. Proceeding from that, an array of social stereotypes is ascribed to different sides of this spectrum, lower culture being connected with low prestige. Given that in this case, linguistic features can be exponents of lower stereotypization, speakers may try to avoid them in certain situations.

Many of these linguistic and extra-linguistic factors can be seen in effect in the Torlak dialectal continuum, as well as in its Timok area, which is the focus of this dissertation. In Timok, there is a multitude of historically, geographically, socially, and linguistically conditioned factors that shape the current linguistic situation.

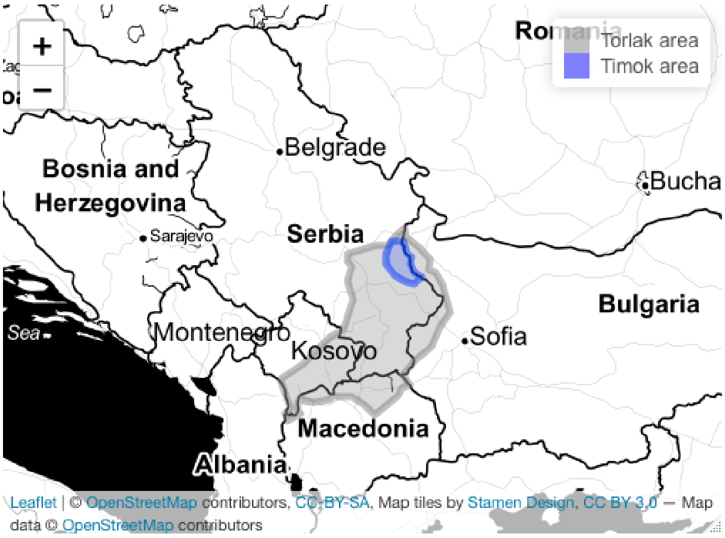
Articles of this dissertation analyze and describe how certain linguistic and extra-linguistic factors, and their interaction, influence linguistic variation in the Timok variety. The following outlines the general circumstances that have been considered in this study. Timok is a transitional South Slavic variety, situated in the contact zone between East South Slavic and West South Slavic linguistic areas, and it intersects grammatical features from these typologically distinct groups. Additionally, Timok has seen demographic changes in recent decades. At the same time, within the context of Serbian varieties, Timok is highly stigmatized and subject to social discrimination, resulting in strong influences from the standard language. This multi-directional overlap of influences makes Timok rich in variation and motivates a contemporary analysis of the dialect. Tracing the interacting driving forces of this variation and the potential changes they bring about extends the relevance of Timok beyond Slavic linguistics, offering broader insights into language variation and change.

In what follows, Section 2 introduces the specific linguistic situation in Timok, outlining the historical, geographical, and sociopolitical factors that have shaped its current state. Section 3 describes the quantitative and computational approaches used to process the Timok data, detailing the challenges and processes involved in data collection, transcription, and annotation. Section 4 presents the variationist theory and methodology used in this research, and the proceeding empirical framework used to analyze linguistic features and their variation. Together, they provide context for the analysis of the Timok dialect and the integration of methodologies employed to uncover its linguistic dynamics.

3. The Contemporary Timok Variety

In this section, we provide a concise overview of the factors that have shaped the current linguistic situation in the Timok region. This context is essential for understanding the unique sociolinguistic dynamics that underpin the research presented in this dissertation. By exploring the historical, geographical, and sociopolitical influences on the Timok dialect, we lay the groundwork for a deeper investigation into its linguistic variations and their implications within the broader framework of Slavic linguistics.[2]

Timok is a distinct variety of the Serbian language, spoken in the Timok region in southeastern Serbia. It is part of the Prizren-Timok dialectal zone of Serbian (Ivić, 1985: 110) and is considered a subset of the broader Torlak dialectal area (Figure 1) (Sobolev, 1998; Friedman, 2006), which straddles the linguistic crossroads between West and East South Slavic languages (West: BKMS, Slovene; East: Macedonian, Bulgarian).



[2] For a more detailed description of the various factors influencing the Torlak area, see Sikimić et al., 2023, Sobolev et al., 2023.

Figure 1: Map of Torlak and Timok area

Geographically, the Timok region is characterized by its relative isolation, compounded by natural barriers such as the Balkan and Svrljiške Mountains, Tresibaba, and the rivers Nišava, Timok, and Grliška. This isolation has helped preserve older linguistic forms and traditions that might have otherwise evolved under external influences. The area stretches from the Serbian-Bulgarian border in the east to the mountainous landscapes that define its northern and southern bounds, encompassing a network of rivers that have historically supported its agrarian communities. The administrative area today encompasses 95 villages, the majority being in the municipality of Knjaževac, while others form part of the municipalities of Zaječar and Svrljig (Dinić, 2008, frontmatter; Ćirković, 2017).

This variety is of significant interest due to its position at the intersection of major Slavic linguistic groups. Timok shares many linguistic features with its Balkan neighbors, making it a part of the well-established Balkan Sprachbund—a linguistic area comprising several Balkan languages that have developed similar features through extensive language contact (Lindstedt, 2000; Joseph, 1992; Friedman, 2006; Asenova, 2002; Sobolev, 2003). This linguistic affiliation is crucial for understanding both the shared and unique aspects of the Timok dialect within the broader context of South Slavic languages.

Various historical influences have affected the Torlak area over the last centuries, thus contributing to the socio-political situation that has shaped the region linguistically until today.[3] Over the last two centuries, this area has experienced considerable shifts in control due to the socio-political turbulence surrounding the Ottoman and Habsburg Empires and the consequential Balkan and World Wars. Timok was incorporated into the Principality of Serbia in 1833 (Milićević, 2012), while parts of this linguistic area, including the Lužnica region south of Timok, were later integrated following the Congress of Berlin in 1878 (Šantić & Martinović, 2007). The more southern reaches remained under Bulgarian jurisdiction until the end of World War I (ibid.). Although the current Serbian-Bulgarian political border has been in place since the end of the 19th century, varying periods of occupation during the 20th century further complicated the dialect's development and influence. Notably, Bulgarian

[3] For a more elaborate overview of the historical context, see Sikimić et al., 2023.

occupation during World War II introduced Bulgarian as the official language in parts of Timok. The splitting of the borders has obstructed communication between linguistically close communities. The Timok region has ended up under the political rule of Serbia, which includes language policies and planning, as well as a cultural landscape.

As a predominantly rural area with many remote villages, Timok hosts a population that still maintains many pre-industrial cultural and religious practices, blending pre-Christian elements with Christian traditions (Zečević, 2008a, b; Jovanović, 2000, 1995). This blend not only adds a layer of richness to the local culture but also provides a unique anthropological backdrop for the study of language variation within the region.

In recent decades, Timok has experienced significant demographic shifts, primarily due to economic changes and urban migration patterns (Penev & Marinković, 2012). The collapse of Yugoslavia led to an economic crisis that affected Serbia, including the Timok region. The post-Yugoslav era has been marked by a decline in industrial activities, especially in places like Knjaževac, once a bustling industrial hub (Mitrović, 2020). The economic downturn resulted in a cultural and technological dichotomy between the lifestyle of the pre-industrial and modern eras, alongside the rise and fall of economic conditions through the last century.

The movement of younger populations towards larger urban centers in search of better employment opportunities and living conditions has led to a depopulation of the more remote villages, leaving an aging population behind (Penev & Marinković, 2012). This demographic change has implications for language use in the region. In Knjaževac and other urban centers, the influence of standard Serbian is more pronounced due to the educational system, public administration, and media, which promote the standard language. Teachers in local schools often discourage the use of dialectal forms, favoring standard Serbian, which further influences the linguistic landscape (e.g., Dragićević, 2001; Brujić, 2016). Conversely, the villages retain more dialectal features, reflecting a more traditional speech pattern that resists standardization.

Additionally, the low prestige of southern dialects, particularly Torlak, has influenced social attitudes towards the dialect. Among the Serbian dialectal continuum, southern dialects hold low prestige, with Torlak being the most depreciated due to noticeable linguistic differences (Petrović, 2015; Ćirković, 2018; Sikimić et al., 2023;

Konior et al., 2023). These dialects are often stigmatized and associated with lower socio-economic status or lower educational levels, which affects how speakers of these dialects perceive themselves and are perceived by others in broader societal contexts. This stigmatization contributes to the endangerment of the dialect, as younger generations may shy away from using it, leading to its gradual decline in daily use. The prestige associated with standard language forms over dialectal variants affects language use patterns among the younger population, especially in formal settings. However, despite these influences, the dialectal forms persist in informal settings and among older community members, reflecting a social stratification based on language use.

Overall, the sociolinguistic landscape of Timok is characterized by a complex interplay of cultural retention and linguistic shift, influenced by economic conditions, migration patterns, educational policies, and social attitudes towards different linguistic forms. As a result of these circumstances, the Torlak variety has been placed on UNESCO’s List of Endangered Languages (Salminen, 2010). This landscape provides fertile ground for sociolinguistic research, offering insights into the mechanisms of language shifts under a multitude of influences.

3.1. Overview of Linguistic Research on the Timok Variety

Before the research outlined in this dissertation commenced, the contemporary understanding of the Timok dialect was notably limited. The first descriptions of the Timok variety within the Serbian dialectal continuum were written by Broch (1903) and Belić (1905), followed by Stanojević (1911). More recent accounts can be found in Bogdanović (1979) and Dinić (2008). Historical approaches to dialectal studies in this region relied heavily on classical dialectological methodologies, often reproducing descriptions across generations without substantial empirical verification or acknowledging the evident linguistic variation (compare Belić, 1905; Stanojević, 1911; Dinić, 2008). Such tendencies resulted in a static depiction of the dialect that insufficiently reflected the dynamic nature of language use within the community.

In recent decades, the socio-economic conditions following the dissolution of Yugoslavia impacted all sectors, including academic research, with dialectology receiving scant attention due to its niche status within the humanities. Additionally, language policy in Serbia has historically prioritized the standard language, sidelining non-standard varieties and influencing the scope and focus of linguistic research. This emphasis on standardization has often overshadowed the rich linguistic diversity present in non-standard dialects like Timok.

The most comprehensive source of linguistic data prior to this study was provided by Andrey N. Sobolev (1998) in his seminal work, which included dialectal maps, distinctive linguistic feature descriptions, and transcripts. This was one of the studies, including the work by Alexander (1975), that placed Timok in a broader linguistic context. While Sobolev's work is impressive, it predominantly focused on non-standard segments spoken by prototypical dialect speakers[4], neglecting variations and the interplay between different linguistic varieties, including the standard form. This selective approach underscored the need for a more inclusive and methodologically diverse research framework.

Given the context outlined in the previous paragraphs, and leveraging quantitative and computational linguistic research approaches, this dissertation aims to address these historical oversights by providing a comprehensive, empirically grounded study of the Timok dialect. By broadening the research scope to include interactions between different linguistic varieties and employing statistical linguistic methodologies, this study seeks to offer a richer, more nuanced understanding of the Timok dialect. The integration of traditional and modern linguistic approaches aims to enhance the dialectological study of Timok, making significant contributions to the fields of Slavic linguistics and sociolinguistics.

3.2. Linguistic variation in Timok

Timok exhibits significant inter- and intra-speaker linguistic variation, influenced by its unique affiliations and historical connections with both the Balkan and West South Slavic languages. Throughout history, Timok has adopted some Balkan innovations such as the post-positive article, subjunctive marking, object reduplication, etc. (Lindstedt, 2000: 287-288; Joseph, 1992: 1-4; Živojinović, 2021). Concurrently, Timok experiences considerable influence from the prestigious standard Serbian, which introduces variability in the use of dialectal versus standard features among its speakers. Detailed descriptions of all the dialectal features and their areal distribution

[4] For the definition of the prototypical dialectal speakers, see Auer 1995: 10.

are available in the existing literature (for the most up-to-date set of features see Sobolev et al., 2023). However, this section will focus only on a select few that are particularly relevant for the research papers included in this dissertation, illustrating the diverse linguistic phenomena that characterize the Timok dialect, with a special focus on their variation.

The linguistic variation in Timok is not only evident in the usage of specific features but also in the degree to which individuals incorporate elements of standard Serbian alongside traditional dialectal forms. It is common for speakers to blend features from Balkan Slavic and standard Serbian, with the most distinctive dialectal features often being subject to conscious modification. This dynamic results in linguistic outputs where many linguistic features exhibit multiple variants within the same community. This dichotomy is evident in many linguistic features, such as the usage of accentuation, with coexisting variants. For instance, certain words may appear with accents in two distinct positions, highlighting a diverse phonetic landscape (Table 1)[5].

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

Additional instances of these co-existing structures are evident in the usage of possessive pronouns (1).

(1) a. *oženil* *se* *sa* *babu* ***moju***
marry.PPART.M.SG REFL with grandma.F.OBL.SG my.F.OBL.SG
'He married my grandmother.' (Timok)[6]

b. *došla* *edna* *vračka* *kod* ***mou***
come.PPART.F.SG one.F.NOM.SG psychic.F.NOM.SG at my.F.OBL.SG

[5] See more in Article 2 of the dissertation.

[6] Unreferenced examples from Timok are extracted from the Spoken Torlak dialect corpus 1.0 (Vuković 2021; 2020)

sveksrvu
mother-in-law.F.OBL.SG
‘A psychic visited my mother-in-law.’ (Timok)

c. *ja* *sam* *si* *kod* *baštu* *bila [,]* *nesam* *kod*
I.NOM AUX REFL at my.F.OBL.SG be.PPART.F.SG AUX.NEG at
ma *mamu*
my.F.OBL.SG mother.M.OBL.SG
‘I was staying at my father’s, not my mother’s.’ (Timok)

d. *od* ***mu*** *decu* *nesam* *patila*
from my.F.OBL.SG grandma.F.OBL.SG AUX.NEG suffer.PPART.F.SG
‘I did not suffer because of my children.’ (Timok)

Moreover, variation is also observed in the presence or absence of certain grammatical markers, such as the post-positive article (2), which appears to be used optionally in different linguistic contexts[7].

(2) *mečka-ta* *dodila* *noćas*
bear.F.NOM.SG-DEM come.PPART.F.SG last night
‘The bear came last night.’ (Timok)

The system of argument structure marking in Timok is another feature illustrating the variety's linguistic variation. Western South Slavic languages, including Serbian, utilize a complex system of inflectional case forms to indicate different syntactic roles within sentences. In contrast, Bulgarian and Macedonian have largely abandoned inflectional case endings in favor of prepositional phrases. For example, in Standard Serbian, the role of a recipient is typically marked using an inflected dative form (3a). Conversely, Bulgarian uses a prepositional phrase with *"na"* followed by the noun in its base form to indicate the recipient (3b). In Timok, the marking of recipients often includes both strategies: it can employ both prepositional and oblique case forms (4a), demonstrating a hybrid approach that retains elements of both West and East South Slavic languages. Additionally, instances where Timok solely uses the inflected form, akin to the standard Serbian model, are also documented (4b, c) (for more examples

[7] For more on the variation in the use of the post-positive article (short demonstrative), see mainly Article 4, but also Article 2 and 3.

see Vuković et al., 2023, Escher & Vuković 2020). This dual method underscores the dialect's position at the linguistic crossroads and its capacity for syntactic variability.

(3) a. *Željko* *[]* *često* *govori* *NE* ***kandidatima**.*
Željko.M.NOM.SG often say.3SG.PRES no candidate.M.DAT.PL
'Željko often says no to the candidates.' (Miličević & Ljubešić 2016)

b. *Pomošt* *se* *dava* ***na*** ***hora***
help REFL give.3SG.PRES on people
'Help is given to the people.' (Erjavec et al. 2021)

(4) a. *Ja* ***na*** ***njenu*** ***ḱerku*** *ponesém*
I.NOM on her.OBL.SG daughter.OBL.SG bring.1SG.PRES
orasi.
walnut.M.ACC.PL
'I bring walnuts to her daughter.'

b. *i* *u* *polako* *u* *sebe* *molitve* *čitam*
and in slowly in myself.OBL prayer.F.ACC.PL read.1SG.PRES
bógu
God.M.DAT.SG
'And slowly to myself I read prayers to God.'

c. *takoj* ***meni*** *pričali*
that way I.DAT read.PPART.M.PL
'That is how they told me.' (Vuković et al, 2023)

Given the multitude of influences it includes, Timok represents a valuable linguistic resource, yet it has historically been under-studied and lacking in resources (Sikimić et al., 2023). The complex interplay of historical, geographical, and socio-cultural factors has crafted a distinctive landscape of linguistic variation, positioning Timok as a dynamic "living laboratory" for exploring the effects of various determinants on linguistic features. This setting offers a unique opportunity to examine the mechanisms of linguistic variation, the interactions between different linguistic features, and the real-time processes of grammaticalization that reflect stages previously documented in the evolution of Slavic and other European languages. Notably, Timok provides insights into the loss of case inflections – a phenomenon extensively observed in Germanic and Romance languages – and the grammaticalization of definite articles (cf. De Mulder & Carlier, 2011; Greenberg, 1978), as well as the Slavic-specific omission of auxiliaries in the perfect tense. Today, the urgency of researching the rich linguistic variation present in the region is underscored by the endangered status of the dialect.

This dissertation contributes significantly by providing valuable empirical data and a methodological framework that supports a thorough understanding of these linguistic phenomena. The initial outputs of this research have already offered significant insights into various dialectal features, enhancing our knowledge of the richness of linguistic variation in Timok. The foundational work for the study in this dissertation began with comprehensive fieldwork conducted from 2015 to 2018 (Ćirković, 2018; Ćirković et al., 2023). The data collected was used for the development of a corpus through meticulous data processing (Section 4; Vuković, 2021; Miličević et al., 2023), which in turn serves as the basis for the linguistic analysis of variation, some aspects of which are detailed in this dissertation.

The subsequent sections present the specific context of Timok, outlining the challenges and approaches taken in this linguistic inquiry. They discuss the data and methodologies employed to investigate the nuances of linguistic variation. This work not only contributes to our understanding of a less-documented linguistic area but also enhances the application of these findings to general linguistic theories and the advancement of NLP for low-resourced languages.

4. Language variation as a challenge and goal for data compilation

The Timok dialect exemplifies common challenges faced in the domains of dialectology, spoken language corpus linguistics, and natural language processing (NLP). The creation of electronic resources for non-standard languages is notably resource-intensive, contributing to the scarcity of such resources compared to those for standard languages. Moreover, the performance of NLP systems often declines when confronted with language variation, a phenomenon well-documented in the literature (Beal et al., 2003; Zampieri et al., 2020; Aepli, 2018: 1). This is partly because spoken dialect corpora demand specific skills for effective data access and processing. Unlike with standard languages, where many tools are readily applicable, analyzing dialects like Torlak requires a meticulous, manual approach to ensure accurate transcription and morphosyntactic analysis, necessitating a profound understanding of the dialect (cf. Scherrer et al., 2019).

Each step in the data collection and processing introduces unique challenges. Despite significant advancements, NLP tools are not universally applicable across languages (Ponti et al., 2019; Bender, 2009, 2011), and standard tools often fail when directly applied to dialects (Zampieri et al., 2020: 597), exacerbated by the inherent variability between dialects. Nevertheless, computational linguistics also offers invaluable tools for linguistic analysis, particularly when faced with large data collections. It can aid language description, and the development and testing of linguistic theories (Bender & Langendoen, 2010; Nerbonne, 2009).

This dissertation employs a methodological framework designed to address these challenges, drawing on existing research and providing resources to support variationist studies. It describes the systematic process of data collection, curation, and analysis necessary for compiling a comprehensive corpus. This work aims to enhance the integration of variable data into language resources and improve processing techniques for non-standard varieties. The initial article in this series outlines the corpus creation process, while the subsequent articles showcase a corpus-based approach to investigate specific linguistic features and their variation across the Timok dialect, as detailed in the research papers included in this dissertation. This comprehensive approach not only tackles the technical difficulties associated with dialect data but also enriches our understanding of linguistic variation and its implications for NLP and dialect studies.

4.1. Fieldwork

The data foundational to this research comprises transcripts from field recordings conducted in the Timok region between 2015 and 2018 (Ćirković, 2018)[8]. These fieldwork sessions provided a direct glimpse into the sociolinguistic dynamics of the area. Notably, there was a marked demographic imbalance: villages primarily housed older populations, while younger residents predominantly lived in urban centers. The impact of Timok's perceived social standing was visible; individuals often self-corrected their speech or expressed embarrassment, highlighting the dialect's stigmatization and the influence of the more prestigious standard Serbian.

[8] Videos collected across various Torlak-speaking locations are available through the team's YouTube Channel Terenska Istraživanja URL: https://www.youtube.com/channel/UC4EpCSAnEb2RIsIRY7pfNdQ. [Accessed on September 13, 2021]

The fieldwork involved recording a diverse group of speakers in terms of age and language production on the spectrum between standard and dialect. Efforts were made to capture authentic linguistic expressions, deliberately seeking those who refrained from adjusting their speech and trying to reduce the observer's paradox. Particularly in rural areas, we targeted speakers who exhibited a predominant use of non-standard dialectal features, striving to record at least one such speaker per location. The semi-structured interview format facilitated this approach, using open-ended questions to elicit extensive, naturalistic responses. Topics were chosen to resonate with interviewees, encouraging them to engage freely and comfortably in their native dialect while researchers minimized their interference or tried to use the dialect for questions. This method not only ensured a rich collection of genuine linguistic data but also supported the goal of documenting the vernacular as faithfully as possible, providing a robust basis for analyzing linguistic variation in Timok.

4.2. Compiling a representative corpus sample

To establish a robust data foundation for analyzing linguistic variation in Timok, a carefully selected subset of audio recordings from fieldwork was transcribed and processed. Processing spoken language, particularly non-standard dialects, is an inherently time-intensive task that requires specialized knowledge due to significant deviations from standard language norms and internal variability.

The challenge lay in devising a sampling methodology that could adequately represent both the highly non-standard dialectal features and the range of variation seen across different locations. Ideally, the sample would need to have balanced demographic representations across gender and age groups. However, the availability of speakers often constrained this ideal setup. To address this, the corpus was structured into several speaker categories, which were internally marked with metadata labels:

- Highly non-standard speakers: This category includes at least one speaker from each selected village, identified by a native expert as a strong representative of non-standard features (referenced from Belić, 1905; Stanojević, 1911; Dinić, 2008). Typically, these were older women.
- Speakers influenced by Standard Serbian: This contrasting group comprises participants who displayed considerable variation towards the standard variety

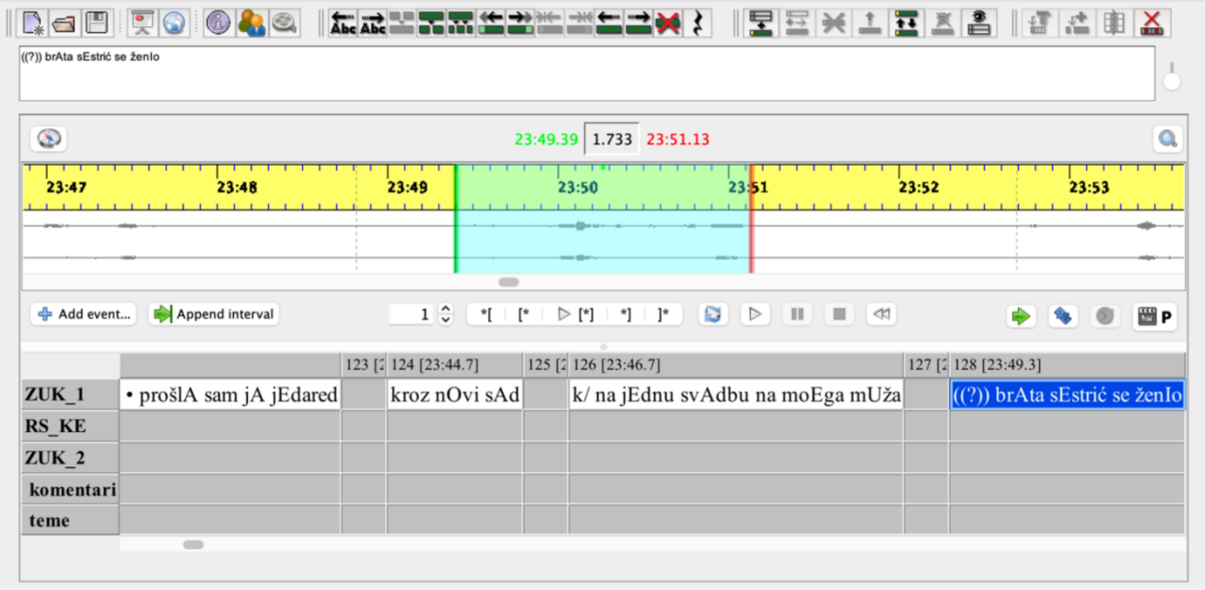
or tended to correct themselves when they spoke. This group consisted of 11 high school students known to exhibit a closer affinity to standard Serbian, allowing for a comparison across age, and participants present in the interviews with the highly non-standard speakers but less active during the interviews, often children or neighbors of the main interviewee.

The corpus ultimately comprises a sample of 500,000 tokens, providing a rich dataset for examining both the highly non-standard Timok dialect variants and the gradations towards standard Serbian. It is important to note that the corpus does not maintain an optimally balanced representation of social demographics. This imbalance reflects the focus of the initial fieldwork on documenting older speakers and more distinctive non-standard forms for preservation purposes. The strategy of prioritizing areal coverage and compiling a large dataset has been effective, offering robust data that supports reliable frequency analyses and allows for comprehensive areal and social comparisons among a broad range of speakers.

The corpus, created through semi-structured interviewing methods and an overarching questionnaire, includes a large array of comparable texts, discussing a number of recurring topics, derived from various speakers across multiple locations. This format provides a robust foundation for diverse comparative analyses. Unlike parallel texts, which are translations that may carry influences from the source language, comparable texts maintain thematic consistency and, to some extent, control linguistic content without imposing another language's structural forms (Scherrer, 2012). This approach offers significant advantages over traditional dialectological methods, such as questionnaires that target specific linguistic features. Historically, dialect surveys have aimed to document either all possible structures or solely the most conservative aspects of a dialect. This traditional method tends to omit less conservative forms that frequently occur in spontaneous speech, often providing a skewed or narrowly focused view that lacks contextual richness essential for understanding the relative frequency and distribution of linguistic features (cf. Nerbonne & Kretzschmar, 2013). In contrast, the use of semi-structured interviews promotes the collection of nearly unrestricted authentic speech, providing a richer and more nuanced understanding of the dialect. Moreover, the corpus of comparable texts allows for controlled diatopic and inter-speaker comparisons. While this methodological approach effectively captures a wide range of linguistic phenomena, it may not adequately represent rare phenomena not specifically elicited during interviews. However, the extensive size of the sample helps mitigate this limitation by ensuring a broad coverage of linguistic variations.

4.3. Transcription

Transcripts were created using the Partitur-Editor from the EXMARaLDA software suite (Schmidt, 2009), designed for transcription and annotation. Figure 2 shows an a segment of a Timok transcript in the EXMARaLDA interface. The semi-orthographic transcription method employed here adheres to standard orthography to approximate phonological variation, a technique successfully implemented in several dialect studies (Anderwald & Wagner, 2014; Johannessen et al., 2014; Vuković & Miličević, 2017; Vuković, 2015). This approach enables the transcription to reflect the complexity of dialectal variation, which is crucial for revealing subtle linguistic differences in the corpus. However, at the same time, this complexity poses challenges for automated processing, as discussed later.



((?)) brAta sEstrić se ženlo

		123 [2	124 [23:44.7]	125 [2	126 [23:46.7]	127 [2	128 [23:49.3]
ZUK_1	• prošlA sam jA jEdared		kroz nOvi sAd		k/ na jEdnu svAdbu na moEga mUža		((?)) brAta sEstrić se ženlo
RS_KE							
ZUK_2							
komentari							
teme							

Figure 2: Transcript segment in *EXMARaLDA*.

An attempt was made to normalize the Timok data to standard forms as a way to reduce this complexity and to simplify annotation and further processing (see Samardžić et al., 2016: 4063). Unfortunately, automating the normalization achieved insufficient accuracy, highlighting the substantial linguistic variation within the Timok dialect and underscoring the difficulties in standardizing such data.

The segmentation of transcripts was implemented to facilitate detailed linguistic analysis. Each speaker was provided with a distinct annotation layer, allowing for focused examination of individual and comparative linguistic patterns. Spoken language was segmented into utterances based on intonation, structural composition, and semantic content. Segmentation into utterances revealed the inherent challenges in processing spoken language due to the less structured nature of speech: firstly, the division into utterances itself, as well as further processing that relies on sentence structure, like parsing techniques and part-of-speech tagging. While spoken utterances do not align perfectly with written sentence structures, they serve as practical units for syntactic analysis within spoken discourse. Transcript texts with annotations are available as TEI XML[9] and as plain text format (Vuković 2021, 2020).

4.4. Annotation

Accurate grammatical annotation is essential for any detailed linguistic analysis and for studies of linguistic variation across large datasets. Annotation of morpho-syntax and lemmas is a standard task for electronic corpora, and a problem that has been solved for most written standard languages. Annotating morpho-syntax and lemmas in electronic corpora is a well-established practice for written standard languages, achieving near-perfect accuracy, especially in languages with low morphological complexity like English. However, non-standard varieties such as the Timok dialect present unique challenges, significantly complicating the training of even basic annotation tools.

One major issue for the automatic processing of dialects is precisely the amount of variation present (Samardžić & Ljubešić, 2021; Zampieri et al., 2020; Aepli & Seinrich, 2021). A significant challenge in processing dialects like Timok is the extensive linguistic variation present (Samardžić & Ljubešić, 2021; Zampieri et al., 2020; Aepli & Seinrich, 2021). Normative varieties typically exhibit consistent morphological markers for grammatical categories, such as the masculine singular past participle in Standard Serbian, which ends in -o, preceded by a vowel (e.g.,

[9] Following TEI guidelines for speech transcription (TEI Consortium 2021)

gledao, ‘looked’). In contrast, Timok displays multiple morphological variants for the same grammatical form. For example, in the corpus, the past participle occurs with two endings, the standard -o, and the dialectal -l, such as in gledao and gledal (‘looked’), respectively. This variability introduces complexity that challenges traditional machine learning algorithms used for linguistic tagging.

Furthermore, the effectiveness of these algorithms often relies on large datasets, which are seldom available for non-standard, low-resource languages (Mikolov et al., 2013; Zhou et al., 2020; Hedderich et al., 2021; Magueresse et al., 2020). In addressing these issues for Timok, low-resourced NLP techniques were used (a subdomain of NLP that deals with processing languages with little data (e.g., Hedderich et al., 2021; Magueresse et al., 2020; Samardžić & Ljubešić, 2021; Aepli & Seinrich, 2021)). More specifically, the Timok tagger training data combined 27,000 manually annotated tokens of dialect material with existing data from standard Serbian due to their similarities (Ljubešić et al., 2016). This approach achieved an 84.61% accuracy rate for part-of-speech tagging and 92.62% for lemmatization, which, while commendable, still falls short of the performance seen with standard varieties. Closer error analysis pointed to a higher error rate in highly variable word forms.

Efforts to enhance the tagger's accuracy by doubling the dialectal training sample did not yield the expected improvements, emphasizing the impact of linguistic variation on NLP tool efficacy. Other methods, such as introducing random orthographic errors to simulate dialectal variation (a technique used successfully with Swiss German, see Aepli & Seinrich, 2021), were ineffective for Timok.[10] This outcome suggests that not all types of linguistic variation are alike and that different dialects may exhibit unique challenges for computational approaches, pointing to the limitations of current NLP technologies when applied to linguistically diverse and variable data.

4.5. Meta-data

[10] Future research could explore alternative strategies such as implementing rule-based semi-automatic normalization to address specific phonological discrepancies between Timok and standard Serbian. This approach would involve creating a normalized intermediary layer informed by detailed dialect descriptions. Additionally, pre-classifying words in the corpus as either Serbian or Timok before annotation might prove beneficial. This method would involve using two separate taggers for each linguistic variant, though it might result in some loss of contextual information critical for tagging accuracy. To date, these methods remain untested and could provide fruitful avenues for enhancing linguistic processing of the Timok dialect.

.

Meta-data plays an indispensable role in analyzing the interplay between extra-linguistic factors and linguistic variation. The corpus used in this research incorporates comprehensive meta-data that includes speaker age and gender, which are critical in studying sociolinguistic variables (as detailed in Article 1). Additionally, individual studies incorporate geographic meta-data to explore spatial linguistic patterns. This includes information such as Euclidean and travel distances between interview sites and significant cultural or administrative landmarks that influence linguistic practices, such as the proximity to the administrative and cultural center of the region, which acts as a hub for the diffusion of standard linguistic features. Furthermore, the distance from the Bulgarian border is considered to gauge the impact of Bulgarian varieties due to geographical closeness. Altitude is also used as an indicator of geographic isolation, expected to correlate with a higher prevalence of non-standard linguistic features. Such geographic meta-data provide a robust framework for dialectometric analysis, allowing for the mapping of areal linguistic trends and their socio-geographical underpinnings (Nerbonne & Kretzschmar, 2013). This approach illustrates the enrichment of corpus analysis through the integration of meta-data, deepening insights into linguistic variation across the Timok region.

5. Variationist Theory and Methodology

In examining examples such as the various argument marking options depicted in example 4, qualitative and intuitive judgments can only provide a descriptive account of the instances observed. These approaches fall short in elucidating the underlying linguistic patterns or sociodemographic trends that may influence the adoption of one system over another. Moreover, they lack the precision required to systematically analyze trends across speakers concerning demographic attributes or the areal spread of linguistic features.

(4)	a.	*Ja*	***na***	***njenu***	***ḱerku***	*ponesém*	*orasi.*
		I.NOM	on	her.OBL.SG	daughter.OBL.SG	bring.1SG.PRES	walnut.M.ACC.PL
		‘I bring walnuts to her daughter.’					

b.	*i*	*u*	*polako*	*u*	*sebe*	*molitve*	*čitam*

and in slowly in myself.OBL prayer.F.ACC.PL read.1SG.PRES
bógu
God.M.DAT.SG
‘And slowly to myself I read prayers to God.’

c. *takoj* ***meni*** *pričali*
that way I.DAT read.PPART.M.PL
‘That is how they told me’ (Vuković et al, 2023)

While isolated or anecdotal instances offer a glimpse into the language usage, they do not allow for a detailed classification of speakers within a larger sample, nor do they enable an exploration of intricate linguistic patterns within or across such samples. Qualitative methods, though insightful for noting dialectal forms or observing differences between speakers who predominantly use Balkan Slavic or Serbian, do not extend to nuanced intra-speaker variations or systematic cross-demographic analysis. Consequently, in contexts such as these, a data-driven quantitative analysis becomes more appropriate and necessary to uncover the subtle dynamics and complex interplays that qualitative approaches alone cannot sufficiently capture.

The empirical framework adopted in this dissertation draws on established variationist methodologies (e.g., Szmrecsanyi, 2013; Tagliamonte, 2006; Chambers et al., 2003; Schmidt, 2010; Kretzschmar, 1996; Jezsensky et al., 2017; Krug & Schlüter, 2013; Nerbonne & Heeringa, 2010; Labov, 1972). It treats language as a dynamic system comprised of interlinked linguistic features, each with its potential expressions – such as synthetic versus analytic marking of argument structure, or distinct strategies for indicating indirect objects. The occurrence probability of linguistic expressions is influenced by the interplay between the structural and combinatorial properties of grammatical categories that are additionally affected by external social or geographic factors. Some examples of language-internal conditions of variation are constraints on specific phonetic clusters, limits on suffix proliferation, or the difference in retention of inflectional marking in pronouns and nouns. Moreover, interactions between features are common and can surpass linguistic domains, as is the case of the relationship between morphological case and word order that exists in some languages, whereby free word order tends to correlate with the marking of morphological case (Comrie, 1989: 91-98).

The constraints and interactions delimit the absolute extent and conditions for variation and mark the boundaries within which social or other extra-linguistic factors can have an effect. Describing the individual patterns and conditions contributes to the description of the totality of variation in a detailed and substantial manner. When describing individual patterns, “circumscribing the variable context is vital, as it defines the loci of variation and categoricity in the grammar” (D’Arcy & Tagliamonte, 2015: 255). The approach and study design applied in this dissertation allow for the precise data-based description of variation within language, updating its current state. This methodology not only allows for detailed descriptions of linguistic variability but also aids in the theoretical exploration of language dynamics (e.g., such as the effects of the case hierarchy (Blake, 1992), or the grammaticalization status and function of the definite article (Greenberg, 1978)).

The analysis in this dissertation is anchored in the surface manifestations of linguistic features and their formal contexts, extracted from corpora that represent natural language usage. The study adopts a bottom-up approach, focusing purely on the manifested linguistic production of speakers without presupposing any influence from external factors such as social group affiliations. This approach strictly utilizes known linguistic features and their assigned variants within the grammatical frameworks of the South Slavic dialect continuum and the broader Balkan contact area, alongside established universal linguistic trends.

Salient linguistic features are central to this analysis, following a methodology common in variationist studies where such features are deemed crucial for understanding linguistic variability (Trudgill, 1986; D’Arcy & Tagliamonte, 2015; Bayley, 2003; Kerswill, 2003). These features, by virtue of their prominence, are highly susceptible to variation and often highlight speakers' awareness of linguistic markedness. The use of distinctly salient features in Timok, as detailed in Articles 2 and 3, serves as a direct indicator of the dialect-standard continuum (cf. Kerswill, 2003: 515), illustrating how these features differentially manifest across the linguistic landscape.

Each chosen feature is scrutinized for its possible manifestations and categorized into variants. This process further involves calculating the frequency and distribution of each variant across the collected samples and situating these within the relevant linguistic context suggested by both South Slavic and broader linguistic scholarship. Subsequent statistical modeling addresses specific research questions, mapping the distribution of variants against linguistic parameters, physical locations, or demographic groups to determine the influence of linguistic, geographical, or social factors on linguistic variation. This empirical strategy allows for a systematic reconstruction of the variation landscape, layer by layer, unveiling the intricate mechanisms of micro-variation and potential cues for their aggregate impact on broader linguistic changes.

6. Overview of the Dissertation

This dissertation employs empirical quantitative methods from corpus linguistics and statistical modeling to explore the web of language-internal and external influences affecting the Timok dialect. By analyzing a corpus of fieldwork interviews from the Timok region, this research examines selected morphosyntactic features that represent the overt formal layer where linguistic variation is most observable. The analysis deconstructs this variation by investigating the linguistic conditions relevant to each feature, aiming to identify triggers and constraints. Beyond the linguistic context, the study considers how social structures and geographical factors influence linguistic variation in Timok.

The dissertation includes several interlinked studies, each contributing to our understanding of linguistic variation in Timok. It utilizes a multidisciplinary blend of corpus and computational linguistics, variationist approaches, dialectometry, and statistical methods, all situated within the context of South Slavic linguistics. This multifaceted analysis not only sheds light on the specific linguistic situation in Timok but also offers broader insights into linguistic variation and change. The methodologies developed are applicable to other linguistic regions experiencing similar contact dynamics, providing a robust framework for testing theoretical models across different languages.

Articles Overview:

1. *Representing Variation in a Spoken Corpus of an Endangered Dialect: The Case of Torlak* (Vuković, 2021)
   This article details the creation of the *Spoken Torlak Dialect Corpus 1.0* (ibid.), a digital resource comprising fieldwork transcripts from the Timok region. The article describes the technical challenges of assembling a spoken dialect

corpus and the solutions developed to address variation and computational challenges in linguistic data processing.

2. *Under the Magnifying Glass: Dimensions of Variation in the Contemporary Timok Variety* (Vuković et al, 2023)
   This study delves into five representative dialectal features, examining their variation under specific linguistic conditions rooted in the South Slavic and Balkan contexts. The article also considers how these features correlate with extralinguistic factors such as gender, age, and geographic proximity to urban centers or linguistic borders.
3. *Degrees of Non-standardness: Feature-based Analysis of Variation in a Torlak Dialect Corpus* (Vuković et al, 2022)
   This article categorizes Timok speakers into clusters reflecting their use of standard versus dialectal features. The study identifies the most indicative features of language use and positions speakers along a continuum from standard to dialectal forms.
4. *A Corpus-based Analysis of the Grammatical Status of Short Demonstratives in Timok Dialect* (Vuković, forthcoming)
   This paper focuses on one of the most distinctive features of the Torlak dialect, *short demonstratives* (a.k.a. the post-positive article). It investigates the usage patterns, referential functions, and syntactic distributions of short demonstratives, assessing their grammatical status between demonstrative pronouns and definite articles.

7. Discussion and Implications

This discussion section delves deeper into the implications of observed linguistic variation in Timok, illustrating how differentiating between speaker groups based on dialectal feature usage can illuminate the complex interplay between various factors. The presupposition that speakers who frequently utilize certain dialectal features are likely to consistently employ others – sets the basis for a stratification of speakers that reflects different stages of language shift from dialect to standard. Inter-speaker variation between dialectal and standard forms is extensively explored in Article 2, which systematically classifies speakers along a dialect-standard continuum based on their usage of five key linguistic features, or more precisely formal dialect or standard manifestations. Feature distribution in across speakers is addressed by analyzing the frequency of each variant within individual speech samples, positioning speakers on a relative scale for each feature, which in turned allowed for a detailed mapping of language use across the community. Cross-correlation between features not only shows more striking co-variation among the more salient dialectal features but also identifies less salient features—such as accent position and auxiliary omission in the perfect—as more predictive of overall linguistic variation. The distribution of features across samples paired with hierarchical clustering is used to categorize speakers in Timok into three ordered groups reflecting a range of variation between dialect and standard. These groups can be used in contrastive analysis within the variety itself as a crucial way of accessing conditions on the choice in variation between parallel variants, serving as an essential tool for examining the factors influencing speaker choices among varying linguistic alternatives.

Timok serves as a linguistic environment that illuminates broader linguistic phenomena, offering valuable cross-linguistic insights. The area showcases ongoing linguistic processes that parallel historical and contemporary developments in the South Slavic dialect continuum, across the Balkan Sprachbund, and in other Indo-European languages.

Certain broadly widespread linguistic features present in Timok have been dealt with in the papers in this dissertation, specifically in Articles 3 and 4. In Article 3, four selected features are decomposed and analyzed by zooming in on patterns of their interaction with specific underlying linguistic contexts.

A South Slavic feature from the verbal domain found in Timok is the non-pronominal use of the reflexive pronoun *si*. Specifically, the usage of the short dative form of the reflexive pronoun *si* in Timok marks a significant divergence from Serbian and aligns more closely with Bulgarian and Macedonian usage. This feature's presence in Timok but absence in Serbian highlights the Balkan areal linguistic influences on Timok and its role in tracing the diffusion of Balkan linguistic features (Kemmer, 1993). Linguistic analysis of the particle *si* in Timok shows the responsiveness of the particle to the morphosyntactic properties of the verb. Its usage varies depending on factors such as the verb's number, person, reflexivity, and animacy. Additionally, it is influenced by the syntactic structure, particularly the word order, and the lexical semantics of the verb (Article 3; Ćirković, 2021).

The phenomenon of auxiliary omission in the perfect tense, which is observed in Timok, aligns with a broader Slavic linguistic trend, exhibiting varying degrees of occurrence across both contemporary and historical Slavic dialects, including those of neighboring South Slavic regions (Meermann, 2015; Meermann & Sonennhauser, 2016; Dickey, 2013; Friedman, 2002). In Timok, this variation is not arbitrary. The tendency to omit auxiliaries in the compound perfect tense correlates strongly with specific types of verbs—namely, those that are intransitive, non-perfective, and non-modal (Article 3; more detailed account in Escher, 2021). Such patterns suggest structured variability affecting the verbal domain in this dialect, which could be studied in the cross-Slavic context.

Like many Indo-European languages, Timok is exhibiting the loss of morphological case distinctions, a process extensively documented across various language families (Kulikov, 2011; Blake, 2004: 175; Iggesen, 2013). This ongoing transformation in Timok provides a real-time example of grammatical simplification that has been observed in historical language evolution. Article 3 analyzes the choice between synthetic/inflectional or analytic/prepositional marking of the indirect object and possessor in Timok. It points to the interaction between form and function in language, showing that the analytic construction depends on the function that the nominal occupies within the sentence, being more commonly used as the indirect object.

Timok showcases the grammaticalization of the definite article from demonstrative pronouns, a globally observed linguistic transition also evident in many European languages and nearby South Slavic languages such as Bulgarian and Macedonian (Lyons, 1999; Dyer, 2005; Mladenova, 2007). This feature received significant attention in this dissertation due to its salience in the dialect and the way its structural properties and their variation effectively represent Timok's transitional nature. The transformation of demonstratives into articles in Timok provides a compelling case study for understanding broader grammaticalization processes within the context of the South Slavic dialect continuum and beyond. Moreover, given its significance in the Balkan context, this case study demonstrates not only the transitional nature of Timok but also holds implications for regional linguistic dynamics and the structural adaptations arising in language contact among the diverse and yet intertwined languages of the Balkans.

In this context, Article 4 examines whether the post-posed demonstrative clitic in Timok truly holds the grammatical status of a definite article, as frequently asserted, and to what extent it can be viewed in line with the fully grammaticalized definite articles found in neighboring Bulgarian and Macedonian. This question is explored using semantic, morphological, syntactic, and pragmatic criteria to differentiate between a demonstrative pronoun, an anaphoric article, and a definite article. The analysis relies on cross-linguistic theories describing the grammaticalization of articles (Greenberg, 1978; Lyons, 1999: 322-340, Dyer 2005). More concretely, the grammatical properties associated with different grammaticalization stages were considered, as well as properties of nominal constituents, in comparison to the descriptions of the analogous morphemes in Bulgarian and Macedonian (Mladenova, 2007; Vulchanova & Vulchanov, 2010; Vladimirovna, 2014).

The theoretical frameworks were operationalized into observable cues within the corpus, enabling the analysis of linguistic variation at a formal level based on text. The direct cues included the morphological form of the post-positive demonstrative, the part of speech of surrounding elements in the noun phrase, and their linear order and morphosyntactic properties. The indirect cues were derived from context or semantics, indicating either deictic or anaphoric references that determine the interpretation of the demonstratives as either article-like or purely demonstrative. These cues were sourced from both the immediate textual context – such as referential and information-structure markings – and from lexical-semantic analyses of the nominal heads involved.

The analysis reveals that in Timok, the *t*-stem of the post-positive demonstrative primarily functions as an anaphoric article, while the other stems, the *v*-stem and *n*-stem, are used deictically, indicating a demonstrative usage. The semantic properties of the nouns they accompany typically involve countable and concrete nouns, aligning more with the identifiability criterion associated with demonstratives rather than the inclusiveness criterion typical of articles. Moreover, when post-positive demonstratives occur within noun phrases that contain modifiers, they only attach to adjectival modifiers. These patterns suggest that they do not fulfill the grammatical role of a definite article, challenging the prevalent classification in the scholarly literature.

Beyond Timok, similar distinctions within the tri-partite post-posed demonstrative class have not been made for other neighboring varieties with their tri-partite form. For instance, in Macedonian, they are jointly classified as articles despite evidence suggesting that *v*- and *n*-stem forms function like demonstratives (Vladimirovna 2014; Topolinjska 2006; Karapojevski 2020). Therefore, the denominations should be revised for the demonstrative uses of the post-posed demonstratives in Timok and potentially other varieties. [11]

In cases like these, theoretically informed analysis of variation provides solid empirical evidence for categorizing functional elements. More specifically, by empirically testing assumptions about definiteness and articles, tendencies that could extend to other (Slavic) languages were revealed (Belaj et al. 2008; Marušič & Žaucer 2006; Trenkić 2004; Yurayong 2020, Becker 2021: 165-185).

Besides the purely linguistic factors, the analyses of variation of the various non-standard features in Timok incorporate socio-demographic and geographic parameters. The analysis of all the features mentioned above show that the non-standard variants are more common in women and with older speakers. Geographic distribution pattern is visible for some of the features, which indirectly points to the importance of socially induced stratification in Timok. These analyses reveal an extra-linguistic pattern of variation, as an inseparable dimension of language. Notably, these patterns cannot be derived from historical language data and require synchronic data with demographic and geographic meta-information.

The value of Timok is that it provides insights into the mentioned linguistic processes and their variation through authentic contemporary spoken language data. While many other languages affected by phenomena such as case loss and grammaticalization of articles are documented through older written texts (e.g., Allen 1997; Wahlström 2015; Mladenova 2007; McFadden 2020; Walters 2004; Manzini, Savoia 2014) controlled by normative writing conventions, these sources often are unable to capture the true vernacular. Moreover, current descriptions can be incomplete, focusing only on specific linguistic aspects, as exemplified by the particle *si* (cf. Arsenijević 2013; Ivanova, Petrova 2017; Milosavljević 2019). The corpus of interviews from Timok offers itself as a valuable resource for the study of language across various dimensions. At the same time, the studies of variation in Timok provide

[11] It is important to point out that the analysis presented here has not been applied to Macedonian, so it is not possible to extend the tendencies observed in Timok. However, the research by Vladimirovna (2014), Topolinjska (2006), and Karapojevski (2020) suggests that the situation with the *v*- and *n*-stem demonstrative post-positions in Macedonian might bear resemblance to that in Timok.

a more comprehensive understanding of various phenomena in a natural linguistic context.

8. Conclusion

The research presented in this dissertation contributes to the domains of South Slavic dialectology, variationist linguistics, dialectometry, sociolinguistics, corpus linguistics, and low-resourced natural language processing. Its primary focus is on representing and analyzing linguistic variation in non-standard spoken language, specifically the Timok dialect of Serbian.

Traditionally, South Slavic dialectology has followed a classical dialectological approach and often outputs descriptions of dialects oriented towards the description or reconstruction of the most conservative state of a dialect. In such an approach, changes in that prototypical state are often seen as a process of decline in the presence of the standard (see Schmidt, 2010: 202-203 for a general claim).

In contrast, the research presented here reconsiders the traditional notion of a dialect as a system of non-standard features. Through quantitative modeling, it not only demonstrates heterogeneous patterns in language but also reveals the underlying mechanisms behind those patterns. Article 2 shows that even within a single variety, such as older speakers of Timok, notable differences in dialectal speech usage exist. Articles 3 and 4 provide in-depth analyses of individual feature variation, demonstrating their dependence on specific linguistic conditions, which can be systematically analyzed to test linguistic theories and model micro-variation. Furthermore, the formal linguistic analysis findings are correlated with external parameters to uncover social and geographic variation patterns, completing a comprehensive dialectometric analysis. This research thus makes a significant contribution to South Slavic linguistics by providing an empirical and quantitative treatment to a lesser-studied variety.

Apart from tackling individual research questions, the research in this dissertation represents a methodology for assessing variation that is deeply based on data and supported by computational tools. Concurrently, it is inspired by variationist work and motivated by linguistic theory in the treatment of linguistic features. The empirical approach employed helps avoid intuition bias, providing a way of conducting fine-grained analysis and deconstruction of conceptually and theoretically complex phenomena using surface forms from spoken data. The results of the analysis lay out the pattern of variation in Timok, adding to current dialect descriptions, uncovering the micro-variation of individual features, and detailing their distribution across geographic and social dimensions. In a broader perspective, the insights regarding feature variation in Timok respond to theoretical assumptions and thus may contribute to cross-linguistic patterns. Separately from the linguistic analyses, the methodology itself can be replicated in other similar contact situations.

An additional outcome of the work conducted on the Timok dialect is the creation of electronic resources, including the corpus of Torlak fieldwork transcripts, as well as automatic processing tools for morpho-syntactic annotation and lemmatization. The production of resources covering small and non-standard varieties has become increasingly valuable in the context of the ever-growing field of language technology. Large collections of manually processed spoken materials serve as valuable inputs for automatic speech recognition models. Beyond the demands of industrial and commercial purposes, resources for non-standard varieties are extremely valuable for linguistic research.

Moreover, the work presented in this dissertation strives to contribute to adding linguistic information on variation to corpora in a structured way, extending beyond common annotations such as PoS or syntactic parsing. This enhancement could be significant for NLP systems, aiding in tasks like classification, tagging, or parsing, especially for non-standard and low-resourced varieties that are abundant in variation. Furthermore, information on variation across speakers and on aspects of variation in individual features or sociolinguistic and geographic information can facilitate cross-pollination between various disciplines and approaches to language. Including information on possible constraints, formal context, and observed variation patterns in language corpora as annotation would create new possibilities for linguistic research. One method of achieving this could be to create corpus platforms that would use user queries to automatically label corpora based on results from specific searches related to linguistic features and make these resources available to other users (cf. Pintzuk, 2019). Additional information on variation layered on top of tokenized text would generate a more accurate representation of language without over-generalizing homogeneity, while bringing scientific rigor and precision (cf. Kretzschmar, 1996: 36).

Finally, the corpus and linguistic descriptions presented here are an effort at language preservation. The Timok variety, as part of the larger Torlak language area, is just one of numerous dialects threatened by the global loss of language diversity in the wave of globalization. Bromham et al. (2021) identify the number of speakers as the biggest predictor of language endangerment, with road density being another significant factor. The population of the area where the Timok dialect is spoken is over 30,000 (Penev & Marinković, 2012), but as the research in this dissertation has shown, not all these people speak the same variety. Given that the most non-standard variety of Timok is spoken by only a small number of elderly, non-migratory speakers, some dialectal features are likely to slowly diminish from the vernacular. The corpus of the Timok variety preserves some of these features for research and as documentation of language and immaterial heritage, while the research in Articles 2, 3, and 4 documents and describes some of the key dialectal features. As some of these linguistic features exemplify broader linguistic processes, the corpus provides resources for wider linguistic research and the possibility to verify whether the principles observed in Timok apply to other languages.

References

Aepli, Noëmi; Clematide, Simon (2018). Parsing Approaches for Swiss German. In: *SwissText 2018*. Winterthur, 12 June 2018 - 13 June 2018.

Aepli, Noëmi; Sennrich, Rico (2021 pre-print). Improving Zero-shot Cross-lingual Transfer between Closely Related Languages by injecting Character-level Noise. ACL.

Allen, Cynthia (1997). Middle English Case Loss and the ‘Creolization’ Hypothesis. *English Language and Linguistics*, vol. 1, no. 1. 63–89. doi:10.1017/S1360674300000368.

Anderwald, Lieselotte; Wagner, Susanne (2014). FRED – The Freiburg English Dialect Corpus: Applying Corpus-Linguistic Research Tools to the Analysis of Dialect Data. In: Joan C. Beal, Karen P. Corrigan, Hermann L. Moisl (eds.): *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*. Basingstoke: Palgrave Macmillan

Anderwald, Lieslotte. (2011). Are non-standard dialects more “natural” than the standard? A test case from English verb morphology. *Journal of Linguistics*, 47(2), 251–274. http://www.jstor.org/stable/41261753.

Anttila, Arto (2002). 8. Variation and Phonological Theory. In: Jack K. Chambers, Peter Trudgill, Natalie Schilling-Estes (Eds). *The Handbook of Language Variation and Change*. Chichester: Wiley-Blackwell. 305-319.

Arsenijević, Boban (2013). Evaluative Reflexions: Evaluative Dative Reflexive in Southeast Serbo-Croatian, in: B. Fernandez, R. Etxepare (eds.) *Variations in Datives: A Microcomparative Perspective,* Oxford: Oxford University Press, 1–21.

Asenova, Petya (2002). *Balkansko ezikoznanie*. Veliko Tărnovo: Faber.

Ash, Sharon (2003). 16. Social Class. In: Jack K. Chambers, Peter Trudgill, Natalie Schilling-Estes (Eds). *The Handbook of Language Variation and Change*. Chichester: Wiley-Blackwell. 305-319.

Auer, Peter (1995). Modelling phonological variation in German. In Werlen, I. (Ed). *Verbale Kommunikation in der Stadt*. Tübingen: Gunter Narr Verlag. 22–37.

Axelsen, Jacob Bock; Manrubia, Susanna. (2014). River density and landscape roughness are universal determinants of linguistic diversity. In: *Proceedings of the Royal Society B*. volume 281, issue 1784. London: The Royal Society.

Bayley, Robert (2003). 5. The Quantitative Paradigm. In: Chambers, Jack K., Trudgill, Peter, Schilling-Estes, Natalie (eds.). *The Handbook of Language Variation and Change. Chichester*: Wiley-Blackwell.

Beal, Joan C.; Corrigan, Karen P.; Moisl, Hermann L. (eds.) (2007). *Creating and Digitizing Language Corpora, Volume 1: Synchronic Databases*. Palgrave.

Becker, Laura (2021). *Articles in the World's Languages*, Berlin, Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110724424.

Belaj, Branimir; Matovac, Darko; Faletar, Goran Tanacković. (2008). Article-like constructions and the definite-indefinite continuum in Croatian. *Folia Linguistica*, vol. 53, no. 1, 2019. 201-231. DOI: https://doi.org/10.1515/flin-2019-2008

Belić, Aleksandar. (1905). *Dijalekti istočne i južne Srbije*. Beograd: Srpska Kraljevska Akademija.

Bender, Emily M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 3(6):1–26.

Bender, Emily M. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction Between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*. 26–32.

Bender, Emily M.; Langendoen, Terence D. (2010). Computational Linguistics in Support of Linguistic Theory. *Linguistic Issues in Language Technology* 3 (February). https://doi.org/10.33011/lilt.v3i.1213.

Blake, Barry (1992). The Case Hierarchy. *La Trobe Working Papers in Linguistics*, 5, 1-6

Blake, Barry (2004). *Case*. Cambridge University Press.

Britain, David (2010). Contact and Dialectology. In: Hickey, Raymond (ed.). *The Handbook of Language Contact*. Wiley-Blackwell.

Bromham, Lindell; Dinnage, Russell; Skirgård, Hedvig; Ritchie, Andrew; Cardillo, Marcel; Meakins, Felicity; Greenhill, Simon; Hua, Xia (2021). Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*.

Brujić, Suzana. (2016). *Timočki đerdan*. Knjaževac: Narodna biblioteka "Njegoš".

Chambers, Jack K. (2003) 14. Patterns of Variation including Change. In: Jack K. Chambers, Peter Trudgill; Natalie Schilling-Estes (Eds). *The Handbook of*

Language Variation and Change. Chichester: Wiley-Blackwell. 265-283 (onlajn izdanje).

Chambers, Jack K.; Trudgill, Peter; Schilling-Estes, Natalie. (eds.). (2003). *The Handbook of Language Variation and Change*. Chichester: Wiley-Blackwell.

Chambers, Jack K.; Trudgill, Peter. (1998) *Dalectology*. Cambridge University Press.

Ćirković, Svetlana. (2021). Nezamenička upotreba enklitičkog oblika dativa refleksivne zamenice si u timočkim govorima: sintaksički obrasci. *Zbornik Matice srpske za filologiju i lingvistiku* LXIV/1. 97-114. Novi Sad: Matica Srpska.

Ćirković, Svetlana. (2019a). Istraživač i sagovornik u kadru. Taktilni gestovi i njihova funkcija u terenskom antropološko-lingvističkom intervjuu. *Ishodišta* 5. 475–492.

Ćirković, Svetlana. (2019b). Ugroženost životne sredine kao element antropološko-lingvističkih intervjua u Knjaževcu i okolini. In: Sibinović, Mikica; Stojadinović, Vladana; Popović Nikolić, Danijela (eds.). *Knjaževački kraj –potencijali, stanje i perspektive razvoja*. Knjaževac: Narodna biblioteka „Njegoš“, Beograd: Srpsko geografsko društvo. 145–157. URL: http://dais.sanu.ac.rs/handle/123456789/5734

Ćirković, Svetlana (2018a). Upotreba gestova u naraciji: sećanje na gajenje i preradu konoplje u Belobreški (Rumuniji). *Ishodišta* 4. 385–400. URL: http://dais.sanu.ac.rs/handle/123456789/5491.

Ćirković, Svetlana. (2018b). Neverbalna komunikacija u antropološko-lingvističkom intervjuu: analiza multimodalnih transkripata narativa na temu gajenja i prerade konoplje. In: Ćirković, Svetlana (ed.). *Timok. Folkloristička i lingvistička terenska istraživanja 2015–2017*. Knjaževac: Narodna biblioteka „Njegoš“, Beograd: Udruženje folklorista Srbije. 219–238. URL: http://dais.sanu.ac.rs/handle/123456789/5492

Ćirković, Svetlana. (2017). Lekovito bilje kao element tradicijske kulture u okolini Knjaževca. In: Karanović, Zoja (ed.). *Gora božurova (biljni svet u tradicionalnoj kulturi Slovena)*. Beograd: Univerzitetska biblioteka „Svetozar Marković“: Udruženje folklorista Srbije. 199–210. URL: http://dais.sanu.ac.rs/handle/123456789/5495

Ćirković, Svetlana; Konior, Daria V.; Sobolev, Andrey N.; Mirić, Mirjana. (2023). Challenges for Torlak data collection. *Zeitschrift für Slavische Philologie*. Volume 79, Issue 1. Heidelberg: Winter Verlag.

Comrie, Bernard (1989). *Language Universals and Linguistic Typology: Syntax and Morphology*. University Of Chicago Press.

D'Arcy, Alexandra, & Tagliamonte, Sali. (2015). Not always variable: Probing the vernacular grammar. *Language Variation and Change*, 27(3), 255-285. doi:10.1017/S0954394515000101

De Mulder, Walter; Carlier, Anne. (2011). The grammaticalization of definite articles. Heine, Bernd; Narrog, Heiko (eds.). *The Oxford Handbook of Grammaticalization*. Oxford University Press.

Dickey, Stephen M. (2013). See, Now They Vanish: Third-Person Perfect Auxiliaries in Old and Middle Czech. *Journal of Slavic Linguistics*, vol. 21, no. 1. Slavica Publishers. pp. 77–121. URL: http://www.jstor.org/stable/24600449.

Dinić, Jakša. 2008. *Timočki dijalekatski rečnik*. Beograd: Institut za srpski jezik.

Dragićević, Radiša (2001). *Samotinja*. Beograd: Nolit.

Dryer, Matthew S. (2005). Definite articles. In: Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 154–157. Oxford: Oxford University Press.

Erjavec, Tomaž; et al., (2021). *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1431.

Escher, Anastasia (2021). Auxiliary Omission in the Perfect Tense in Timok. *Balkanistica*, 34. The University of Mississippi.

Escher, Anastasia; Vuković, Teodora (2020). Dative clitics as adnominal possessives in the transitional varieties of Southeastern Serbia and Western Bulgaria. *Slověne*. Moscow: RAN.

Friedman, Victor A. (2002). *Macedonian*. (Languages of the World / Materials 117). Munich: LinCom Europa.

Friedman, Victor A. (2006). Balkans as a Linguistic Area. Balkans as a Linguistic Area. In: Brown, Keith, (Editor-in-Chief). *Encyclopedia of Language & Linguistics*, Second Edition, Volume 1. Oxford: Elsevier. 657- 672.

Giles, H., & Ogay, T. (2007). Communication Accommodation Theory. In B. B. Whaley & W. Samter (Eds.), Explaining communication : Contemporary theories and exemplars (pp. 293-310). Mahwah, NJ: Lawrence Erlbaum.

Gooskens, Charlotte. (2005). Travel Time as a Predictor of Linguistic Distance. In: Astrid van Nahl (ur.): *Dialectologia et Geolinguistica* 13(13). Berlin, Boston: De Gruyter Mouton. 1-25.

Greenberg, Joseph H. (1978). How does a language acquire gender markers?. In: Greenberg, Joseph H.; Ferguson, Charles A.; Moravcsik, Edith A. (eds.). *Universals of human language*, vol. 3. Stanford, CA: Stanford University Press. 47–82.

Guy, Gregory R. (1991). Explanation in variable phonology. Language Variation and Change (3): 1 22.

Guy, Gregory R. (1994). The phonology of variation. In: Beals et al. (eds.), CLS 30, Volume 2: The Parasession on Variation in Linguistic Theory, Chicago: CLS. 133 49.

Hedderich, Michael A.; Lange, Lukas; Adel, Heike; Strötgen, Jannik; Klakow, Dietrich. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In: *Proceedings of the NAACL-HLT 2021*. ACL

Hickey, Raymond (ed.). (2013). *Handbook of Language Contact*. Chichester (West Sussex): Wiley-Blackwell.

Ivanova, Elena Jurovna; Petrova, Galina Mihajlova (2017). Bolgarskie vozvratye klitiki SE i SI: Omonimija, polisemija, sintaksis, Voprosy jazykoznanija 1/ 2017, 74–104.

Jeszenszky, Péter; Stoeckle, Philipp; Glaser, Elvira; Weibel, Rober. (2017). Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German. *Journal of Linguistic Geography*. Cambridge University Press. 1-23.

Johannessen, Janne Bondi; Vangsnes, Øystein Alexander; Priestley, Joel; Hagen, Kristin (2014). *A multilingual speech corpus of North-Germanic languages*. In: Raso, Tommaso; Ribeiro, Heliana De Mello (eds.): Spoken Corpora and Linguistic Studies. Amsterdam: John Benjamins. 69-83.

Joseph, Brian (1992). The Balkan Languages. In: Bright, W. (Ed.) *International Encyclopedia of Linguistics*. Oxford: Oxford University Press. 153-155.

Jovanović, Bojan (1995). *Magija srpskih obreda*. Novi Sad: Svetovi.

Jovanović, Bojan (2000). *Duh paganskog nasleđa*. Novi Sad: Svetovi.

Kautzsch, Alexander (2000). *The historical evolution of earlier African American English: A comparison of written sources*. Ph.D. dissertation. University of Regensburg.

Kerswill, Paul. (2003). 26. Koineization and Accommodation. In: Chambers, Jack K., Trudgill, Peter, Schilling-Estes, Natalie (eds.). *The Handbook of Language Variation and Change. Chichester*: Wiley-Blackwell.

Konior, Daria V.; Sobolev, Andrey N.; Ćirković, Svetlana; Mirić, Mirjana. (2023). Torlak: Research approaches, the sociolinguistic situation and perception at the beginning of the 21st century. *Zeitschrift für Slavische Philologie*. Heidelberg: Winter Verlag.

Kretzschmar, William A. (1996). Quantitative areal analysis of dialect features. *Language Variation and Change*, *8*(1), 13–39. http://doi.org/10.1017/S0954394500001058

Kroch, Anthony. (1978). Toward a theory of social dialect variation. *Language in Society*, 7(1), 17-36. DOI:10.1017/S0047404500005315

Krug, Manfred; Schlüter, Julia (Eds.). (2013). *Research Methods in Language Variation and Change*. Cambridge: Cambridge University Press.

Kulikov, Leonid (2011). The Proto-Indo-European Case System and Its Reflexes in a Diachronic Typological Perspective: Evidence for the Linguistic Prehistory of Eurasia. *Rivista Degli Studi Orientali*, 84(1/4), 289–309. URL: http://www.jstor.org/stable/43927273.

Labov, William (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Labov, William (1994). Principles of Linguistic Change: Internal Factors . Oxford, UK, and Cambridge, USA: Blackwell.

Labov, William (1997). Resyllabification. In: Frans Hinskens, Roeland Van Hout and Leo Wetzels (eds.), Variation, Change and Phonological Theory , Amsterdam, Philadelphia: John Benjamins Publishing Company. 145 79.

Labov, William, Paul Cohen, Clarence Robins, John Lewis (1968). A Study of the Nonstandard English of Negro and Puerto Rican Speakers of New York City. *Cooperative Research Report 3288*, Vols I and II. Philadelphia: U.S. Regional Survey.

Leopold, Werner (1930). Polarity in language. Curme volume of linguistics studies. Baltimore: Waverly Press.

Lindstedt, Jouko (2000). Linguistic Balkanization: Contact-induced Change by Mutual Reinforcement. *Studies in Slavic and General Linguistics* Vol. 28, *Languages in Contact*: 231–246.

Ljubešić, Nikola; Klubička, Filip; Agić, Željko; Jazbec; Ivo-Pavao. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: ELRA.

Lyons, Christopher (1999). *Definiteness*. Cambridge: Cambridge University Press.

Magueresse, Alexandre; Carles, Vincent; Heetderks, Evan. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. In: *Computing Research Repository (CoRR)*. URL: https://arxiv.org/pdf/2006.07264.pdf. Accessed on 07.12.2021.

Manzini, M. Rita; Savoia, Leonardo M. (2014). From Latin to Romance: case loss and preservation in pronominal systems. *Probus*, vol. 26, no. 2,. 217-248. https://doi.org/10.1515/probus-2014-0008

Marušič, Franc; Žaucer, Rok (2006). he ‘definite article’ *TA* in colloquial Slovenian. *Formal Approaches to Slavic Linguistics* (Vol. 14, pp. 189-204).

Matras, Yaron (2009). *Language Contact*. Cambridge University Press.

McFadden, Thomas (2020). Case in Germanic. In: Putnam, Michael T.; Page, Richard B. (Eds.). *The Cambridge Handbook of Germanic Linguistics*. 282–312. Cambridge University Press.

Meermann, Anastasia (2015). Truncated perfect in Serbian: A marker of distance? In: Sonnenhauser, Barbara & Anastasia Meermann (eds.). *Distance in language. Grounding a metaphor*. Newcastle upon Tyne: Cambridge Scholars Publishing, 95-116.

Meermann, Anastasia; Sonnenhauser, Barbara (2016). Das Perfekt im Serbischen zwischen Slavischer und Balkanslavischer Entwicklung. In: Bazhutkina, Alena & Barbara Sonnenhauser (eds.), *Linguistische Beiträge zur Slavistik. XXII. JungslavistInnen-Treffen in München. 12. bis 14. September 2013*. Leipzig: Biblion Media, 83–110.

Mikolov, Tomas; Kai, Chen; Greg, Corrado; Dean, Jeffrey. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Miličević Petrović, Maja; Vuković, Teodora; Mirić, Mirjana; Konior, Daria; Makarova, Anastasia. (2023). Language Documentation II: "Towards a sociolinguistic corpus of Torlak: Challenges for data processing. *Zeitschrift für Slavische Philologie*. Heidelberg: Winter Verlag.

Miličević, Maja; Ljubešić, Nikola (2016). Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina* 2.0 4(2)

Milićević, M. (2012). *Knjaževački okrug*. Knjaževac: Narodna biblioteka „Njegoš” (reprint of the book from 1876).

Milosavljević, Stefan (2019). Semantika i pragmatika evaluativnog refleksivnog dativa u govoru jablaničkog kraja (u svetlu srodnih fenomena u drugim jezicima/ govorima), in: M. Kovačević, J. Petković (eds.) *Savremena proučavanja jezika i književnosti* X/1. 45–56.

Mirić, Mirjana. (2017). Stepen gramatikalizacije futura prvog u timočkim govorima. *Zbornik Matice srpske za filologiju i lingvistiku* LX/1. Novi Sad: Matica srpska. URL: http://dais.sanu.ac.rs/handle/123456789/4627

Mirić, Mirjana. (2018a) “Gramatikalizacija futura prvog i izostavljanje subjunktivnog markera da u lužničkom govoru južnog tipa zone I (Bukovik)“. Zbornik Matice srpske za filologiju i lingvistiku 61/2, 89–125. URL: http://dais.sanu.ac.rs/handle/123456789/4679

Mirić, Mirjana. (2018b). Upotreba/izostavljanje subjunktivnog markera "da" u konstrukciji futura prvog u timočkim govorima. In: Ćirković, Svetlana. (Ed), *Timok. Folkloristička i lingvistička terenska istraživanja 2015–2017*. Knjaževac: Narodna biblioteka “Njegoš”. URL: http://dais.sanu.ac.rs/handle/123456789/4598

Mirić, Mirjana; Miličević Petrović, Maja; Ćirković, Svetlana. (2021). Digitalizacija jezika i kulture kroz elektronske korpuse: primer timočkih govora. In: Vraneš, Aleksandra (ed.). *Digitalna humanistika i slovensko kulturno nasleđe I (Međunarodna naučna konferencija. Zbornik radova)*. Beograd: Savez slavističkih društava Srbije, Filološki fakultet Univerziteta u Beogradu. 75–94.

Mitrović, Nemanja. (2020). Srbija, Jugoslavija i industrija: Kako se živelo u Zaječaru u vreme industrijskih džinova. *BBC News na Srpskom*, official website. URL: https://www.bbc.com/serbian/lat/srbija-53055717. [Accessed on 19.12.2021.]

Mladenova, Olga. (2007). *Definiteness in Bulgarian. Modelling the Processes of Language Change*. De Gruyter Mouton.

Nerbonne, John (2009). Data-driven dialectology. *Language and Linguistics Compass*, 3 (1). 175–198.

Nerbonne, John; Kretzschmar, William A. (2013). Dialectometry++. In: *Literary and Linguistic Computing*, 28(1). 2-12. DOI: https://doi.org/10.1093/llc/fqs062.

Nerbonne, John; Heeringa, Wilbert (2010). Measuring dialect differences. In Schmidt, Jürgen Erich; Auer, Peter (eds.). *Language and space: theories and methods*. Berlin: Mouton de Gruyter. 550–567.

Nichols, Johanna (2015). Types of spread zones, Open and closed, horizontal and vertical. In: Rik de Busser, Randy J. LaPolla: *Language structure and environment: social, cultural, and natural factors*. Amsterdam: Benjamins. 261–286.

Palander, Marjatta; Helka Riionheimo; Vesa Koivisto (2018). Introduction: Creating and Crossing Linguistic Borders. In: Palander, Marjatta; Helka Riionheimo; Vesa Koivisto (Eds.). *On the Border of Language and Dialect*. Helsinki: Finnish Literature Society. DOI: https://doi.org/10.21435/sflin.21

Penev, Goran; Marinković, Ivan (2012). Prvi rezultati popisa stanovništva 2011. S posebnim osvrtom na promenu broja stanovnika jugoistočne Srbije. In: *Stanovništvo jugoistočne Srbije: uticaj demografskih promena u jugoistočnoj Srbiji na društveni razvoj i bezbednost*. Niš: Centar za naičnoistraživački rad SAU i Univerziteta u Nišu.

Petrović, Tanja. (2015). *Srbija i njen jug*. Beograd: Fabrika knjiga.

Pintzuk, Susan (2019). Adding Linguistic Information to Parsed Corpora. *Linguistic Issues in Language Technology*, vol. 18, no. 1. DOI: https://doi.org/10.33011/lilt.v18i.1435.

Ponti, Edoardo Maria; O'Horan, Helen; Berzak, Yevgeni; Vulić, Ivan; Reichart, Roi; Poibeau, Thierry; Shutova, Ekaterina; Korhonen, Anna (2019). Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics* 2019; 45 (3): 559–601. doi: https://doi.org/10.1162/coli_a_00357

Salminen, Tapani (2010). Europe and the Caucasus. In: Moseley, Christopher (Ed.). *Atlas of the world's languages in danger*. Paris: UNESCO. 32-42.

Samardžić, T. and N. Ljubešić (2021). Data Collection and Representation for Similar Languages, Varieties and Dialects. In M. Zampieri and P. Nakov (eds.) *Similar Languages, Varieties, and Dialects: A Computational Perspective, Studies in Natural Language Processing.* Cambridge University Press

Samardžić, Tanja; Scherrer, Yves; Glaser, Elvira (2016) ArchiMob - A corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

Santa Ana, Otto (1992). Locating the linguistic cycle in vernacular speech: Chicano English and the Exponential Hypothesis. In J. M. Denton, G. P. Chan and C. P. Canakis (eds.), CLS 28: Papers from the 28th Regional Meeting of the Chicago Linguistic Society, Vol. 2: The cycle in linguistic theory. Chicago: CLS. 277 87.

Šantić, Danica; Martinović, Marija. (2007). Naselja Lužnice – geografsko-istorijska i prostorno-demografska transformacija. *Glasnik Srpskog geografskog društva* LXXXVII (2): 115–124.

Scherrer, Yves; Samardžić, Tanja; Glaser, Elvira. (2019). Digitising Swiss German – how to process and study a polycentric spoken language. *Language Resources and Evaluation* 53(4), 735–769.

Scherrer, Yves. (2012). Recovering dialect geography from an unaligned comparable corpus. In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Avignon: Association for Computational Linguistics. URL: https://aclanthology.org/W12-0210. 63-71.

Schilling-Estes, Natalie (2003). Fieldwork: Introduction. In: Chambers, Jack K., Trudgill, Peter, Schilling-Estes, Natalie (eds.). *The Handbook of Language Variation and Change. Chichester*: Wiley-Blackwell.

Schmidt, Thomas (2009). Creating and Working with Spoken Language Corpora in EXMARaLDA. In: *LULCL II: Lesser Used Languages & Computer Linguistics II*.151-164.

Schmidt, Jürgen Erich (2010). 12. Language and space: The linguistic dynamics approach. In: Auer, Peter; Schmidt, Jürgen Erich (eds.). *Volume 1 Theories and Methods: An International Handbook of Linguistic Variation* (pp. 201-225). Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110220278.201

Schneider, Edgar W. (2003). 3. Investigating Variation and Change in Written Documents. In: Chambers, Jack K., Trudgill, Peter, Schilling-Estes, Natalie (eds.). *The Handbook of Language Variation and Change. Chichester*: Wiley-Blackwell.

Sikimić, Biljana, Sobolev, Andrey N., Sonnenhauser, Barbara. (2023). Introduction: (Dis-)entangling traditions in the Central Balkans: Performance and perception. The case of Torlak. *Zeitschrift für Slavische Philologie*. Heidelberg: Winter Verlag.

Sobolev, Andrej (1998). *Sprachatlas Ostserbiens und Westbulgariens*. München: Biblion Verlag.

Sobolev, Andrej (2003). *Malyj dialektologicheskij atlas balkanskih jazykov. Probnyj vypusk*. München: Biblion Verlag.

Sobolev, Andrey N.; Mirić, Mirjana; Konior, Daria V.; Ćirković, Svetlana. (2023). Torlak: Areal Embedding and Linguistic Characteristics. *Zeitschrift für Slavische Philologie*. Heidelberg: Winter Verlag.

Stanojević, Marinko. 1911. Severno-timočki dijalekat. *Srpski dijalektološki zbornik* II: 360–463.

Szmrecsanyi, Benedikt (2013). *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge University Press.

Tagliamonte, Sali (2006). *Analysing Sociolinguistic Variation*. Cambridge University Press.

Tagliamonte, Sali (2011). *Variationist Sociolinguistics: Change, observation, interpretation*. Wiley-Blackwell.

Tauli, Valter. (1958). The structural tendencies of languages. Helsinki: Suomalainen Tiedeakatemia.

TEI Consortium (eds.). (2021). 8 Transcriptions of Speech. In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.3.0., released on 31.08.2021. URL. http://www.tei-c.org/Guidelines/P5/. (Accessed on 9.12.2021).

Terenska istraživanja. (2021). Youtube Channel. URL: https://www.youtube.com/channel/UC4EpCSAnEb2RIsIRY7pfNdQ. [Accessed on September 13, 2021]

Trenkić, Danijela. (2004). Definiteness in Serbian/Croatian/Bosnian and some implications for the general structure of the nominal phrase. *Lingua*, Volume 114, Issue 11. 1401-1427. DOI: https://doi.org/10.1016/j.lingua.2003.09.005.

Trudgill, Peter. (1986). *Dialects in Contact*. Oxford: Blackwell.

Vladimirovna, Boronnikova Natalija. (2014). Status trojnogo člena v makedonskom jazike. In: *Filologičeskie nauki. Voprosj teorii i praktiki*. Gramota.

Vuković, Teodora; Escher, Anastasia; Sonnenhauser, Barbara (2022). Degrees of non-standardness. Feature-based analysis of variation in a Torlak dialect corpus. *International Journal of Corpus Linguistics*. John Benjamins.

Vuković, Teodora; Miličević, Maja (2017). Creation and Some Ideas for Classroom Use of an Electronic Corpus of the Dialect of Bunjevci. In: Filipović, Jelena; Vučo, Julijana (eds.). *Minority Languages in Education and Language Learning: Challenges and New Perspectives*. Edicija Filološka istraživanja danas, Vol 7. Belgrade: Faculty for Philology, University of Belgrade.

Vuković, Teodora; Mirić, Mirjana; Escher, Anastasia; Ćirković, Svetlana; Miličević Petrović, Maja; Sobolev, Andrey; Sonnenhauser, Barbara. (2023). Under the magnifying glass. Dimensions of variation in the contemporary Timok variety.

Vuković, Teodora; Samardžić, Tanja. (2018). Prostorna raspodela frekvencije postpozitivnog člana u timočkom govoru. In: Ćirković, Svetlana. (Ed), *Timok. Folkloristička i lingvistička terenska istraživanja 2015–2017*. Knjaževac: Narodna biblioteka “Njegoš”. 181–200.

Vuković, Teodora. (2015). *Izrada modela dijalekatskog korpusa bunjevačkog govora*. Unpublished Master’s Thesis. Belgrade: Faculty for Philology, University of Belgrade.

Vuković, Teodora. (2019). *Torlak ReLDI Tagger 2019* (GitHub repository). URL: https://github.com/bravethea/Torlak-ReLDI-Tagger-2019. Accessed on 06.01.2022.

Vuković, Teodora. (2020). Spoken Torlak dialect corpus 1.0 (transcription). Slovenian language resource repository CLARIN.SI. URL: http://hdl.handle.net/11356/1281.

Vuković, Teodora. (2021). Representing variation in a spoken corpus of an endangered dialect: the case of Torlak. In: *Language Resources & Evaluation 55*. 731–756. DOI: https://doi.org/10.1007/s10579-020-09522-4.

Vuković, Teodora. (forthcoming). A Corpus-based Analysis of the Grammatical Status of Short Demonstratives in Timok Dialect. *Journal of Slavic Linguistics*. Slavic Linguistics Society.

Vulchanova, Mila Dimitrova; Vulchanov, Valentin. (2010). An article on the rise. Contact-induced change and the rise and fall of N-to-D movement. In. Anne Breitbarth, Christopher Lucas, Sheila Watts and David Willis (Eds.). *Continuity and Change in Grammar*. John Benjamins.

Wahlström, Max (2015). *The loss of case inflection in Bulgarian and Macedonian*. *Slavica Helsingiensia* 47. Helsinki: University of Helsinki.

Walters, Denise E. (2004). Loss and Consequence: An Examination of the Old English Case Marking System As Opposed to that of Other Old Germanic Languages. *Colorado Research in Linguistics*, *17*. https://doi.org/10.25810/k43m-zz07

Wardhaugh, Ronald; Fuller, Janet M. (2014). *An Introduction to Sociolinguistics*. Wiley-Blackwell.

Weinreich, Uriel; Labov, William; Herzog, Marvin. (1968). Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkiel (eds.). *Directions for historical linguistics*. Austin: University of Texas Press. 95–188.

Yurayong, Chingduang (2020). *Postposed demonstratives in Finnic and North Russian dialects*. Helsinki: University of Helsinki. Doctoral Thesis.

Zampieri, Marcos; Nakov, Preslav; Scherrer, Yves. (2020). Natural language processing for similar languages, varieties, and dialects: A survey. Natural Language Engineering, 26(6), 595-612. doi: 10.1017/S1351324920000492

Zečević, Slobodan (2008a). *Mitska bića srpskih predanja*. Beograd: Službeni glasnik.

Zečević, Slobodan (2008b). *Srpska etnomitologija*. Beograd: Službeni glasnik.

Zhou, Ming; Duan, Nan; Liu, Shujie; Shum, Heung-Yeung (2020). Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering*, Volume 6, Issue 3. 275-290. https://doi.org/10.1016/j.eng.2019.12.014.

Zipf, G. K. (1949). Human behavior and the principle of least effort. Cambridge: Addison-Wesley Press.

Živojinović, Jelena. (2021) Torlak clitic doubling: A cross-linguistic comparison. In: *Advances in formal Slavic linguistics 2018*. Language Science Press.

Article 1

Representing variation in a spoken corpus of an endangered dialect: the case of Torlak

by

Teodora Vuković

Abstract

The paper presents a spoken corpus of the endangered Torlak dialect from the Timok area of Southeast Serbia. This dialect expresses a great deal of variation in the use of non-standard features under the influence of standard Serbian (SSr). Accounting for this variation, a specific methodology has been selected for collection, sampling, transcription and annotation. Between 2015 and 2017, semi-structured interviews were conducted in the field eliciting spontaneous speech in the form of long narratives about traditional culture and history. The corpus comprises 500,697 tokens of semi-orthographic transcripts representing 80 h of recording from locations evenly distributed across the Timok area of the Torlak dialect zone, thus enabling a spatial contrastive analysis. The majority of speakers in the corpus are older people whose language represents the highly non-standard variety. In order to allow for analysis of language change under the influence of SSr, the corpus includes a number of younger people whose speech is closer to SSr. Tools for automatic PoS annotation and lemmatization that were lacking were developed based on the existing resources for SSr. For tagger training, a dialect sample of 27,000 manually verified tokens was merged with an existing training set for SSr.

This article was originally published in:

Teodora Vuković, (2021). Representing variation in a spoken corpus of an endangered dialect: the case of Torlak. *Language Resources & Evaluation,* 55, pp. 731-756. DOI: https://doi.org/10.1007/s10579-020-09522-4

Visit the publisher’s page to read the full article.

Article 2

Under the magnifying glass. Dimensions of variation in the contemporary Timok variety

by

Teodora Vuković, Mirjana Mirić; Anastasia Escher, Svetlana Ćirković, Maja Miličević Petrović, Andrey N. Sobolev, Barbara Sonnenhauser

Abstract

The paper focuses on linguistic variation encountered in the contemporary Timok variety in Southeast Serbia. The data collected from speakers across the Timok region reveals rich variation in how linguistic features are used. It Displays overlapping patterns which belong to different registers and varieties, which arises from the contact between West South Slavic varieties and East or Balkan South Slavic varieties. In the analysis, the focus is on the interaction between patterns attributed to standard Serbian (currently major western South Slavic influence and the more typical Balkan manifestations of the dialect. The features analysed are: marking of indirect object and possessor, post-positive demonstratives, dative reflexive si as a particle and auxiliary omission in the perfect tense, all considered to be relative for the distinction analysed. In the first part, the analysis takes into account linguistic factors focusing mainly on the morphosyntactic domain, to reveal what linguistic structures obstruct or facilitate the use of certain forms. In the second part, the four features are correlated with geographic and social parameters. In order to find potential areal patterns of horizontal feature diffusion or study the effect of terrain shape (altitude) or physical distances between locations (the distance from the administrative centre). Regarding social factors, age and gender are correlated with the linguistic information to analyse whether there are differences between men and women, or older and younger speakers in their dialect usage.

This article was originally published in:

Teodora Vuković, Mirjana Mirić; Anastasia Escher, Svetlana Ćirković, Maja Miličević Petrović, Andrey N. Sobolev, Barbara Sonnenhauser, (2023). Under the magnifying glass. Dimensions of variation in the contemporary Timok variety. *Zeitschrift für Slavische Philologie*, Issue 79/1. URL: https://zsph.winter-verlag.de/article/ZSPH/2023/1/9

Visit the publisher’s page to read the full article.

Article 3

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

by

Teodora Vuković, Anastasia Escher, Barbara Sonnenhauser

Abstract

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.

This article was originally 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*. DOI: https://doi.org/10.1075/ijcl.20014.vuk.

Visit the publisher’s page to read the full article.

A Corpus-Based Analysis of the Grammatical Status of Short Demonstratives in the Timok Dialect

Teodora Vuković

Abstract: The present study addresses the question of the status of demonstrative enclitics (short demonstratives (SDs)) in Timok in the process of their grammaticalization from a demonstrative into a definite article. It uses insights from neighboring Bulgarian and Macedonian varieties where this process of grammatical change has resulted in a fully grammaticalized definite article. Different linguistic criteria are used to situate the Timok SD on the grammaticalization scale between a demonstrative, anaphoric article and a definite article. It analyzes the type of referential marking of the three demonstratives (*ovaj, taj, onaj* 'this, that, yonder'; *t-*, *v-*, *n-*forms, respectively), as well as their distribution in noun phrases and the type of noun they select. All findings point to their status as anaphoric articles. However, when it comes to the type of reference, although there is variation, the *t*-form of the SD is dominantly used for anaphoric referencing, while *v*-form and *n*-form are more commonly used deictically. Insight into idiolects reveals that some speakers show a more advanced use of SDs on the grammaticalization scale than others, by using SDs more frequently and exhibiting a more anaphoric use. They tend to select countable and concrete nouns, linking SDs to the deictic meaning of the demonstrative. Within a nominal expression, SD attaches almost exclusively to adjectival modifiers, which suggests that it does not have the status of a functional element marking definiteness.

1. Introduction

Postpositive articles are considered to be one of the typical features of the South Slavic languages associated with the Balkan Sprachbund—Bulgarian, Macedonian, and Torlak (Lindstedt 2000; Friedman 2006)—setting them apart from other Slavic languages, which are typically article-less. Postpositive articles are always identified as one of the characteristics of southeastern Serbian Torlak varieties of Timok and Lužnica (Belić 1905; Ivić 1985), often considered to be their "most important feature" (Ivić 1985: 116–17; Belić 1905: 442). These articles are thus regarded as a salient trait that separates the Torlak varieties from other Serbian dialects and that approximates them to Bulgarian and Macedonian varieties.

Journal of Slavic Linguistics 31(1–2): xx–xx, 2023.

The postpositive article is an enclitic originating from a demonstrative pronoun that attaches to the end of its nominal host.[1] It typically takes the second position in a nominal expression, attaching to the left-most element of the NP, a noun, or a noun modifier. In Bulgarian and Macedonian, these articles act as a marker of definiteness, performing the function of the definite article (Tomić 2006: 49; Stojanov 1983: 115; Koneski 1967).

The development of the definite article in South Slavic languages is attributed to the contact between other Balkan languages, which together constitute the Balkan Sprachbund, sharing several common features, the article among them (Joseph 1992). The definite article in Bulgarian and Macedonian results from a grammaticalization of adnominal demonstrative pronouns (ADPs; Mladenova 2007) that evolved into the cliticized article that we find in contemporary varieties. Grammaticalization involved changes across several linguistic domains. A standalone accentuated pronoun gained another function in its accentless and cliticized form, attaching to the left of a nominal host. The deictic meaning of the ADP expanded to an anaphoric marker and finally to a marker of definiteness (Mladenova 2007). Syntactically, the definite article is a determiner that appears in the left periphery of the NP, which is typical for functional words such as articles in these South Slavic languages (Dimitrova-Vulchanova and Vulchanov 2010, 2011). The demonstrative clitic used in the postpositive position and carrying anaphoric and definite marking has seen an increase in frequency over time and has become an essential element of the Bulgarian and Macedonian NP (Mladenova 2007).

The Timok and Lužnica varieties belong to the periphery of the Balkan Sprachbund. While they do use postpositive demonstrative clitics, they do so much less frequently than standard Bulgarian and Macedonian and also display considerable inter- and intraspeaker variation. Historically, the western Balkan Slavic periphery is known to display fewer postpositive demonstratives; their distribution reveals that they are not fully grammaticalized into markers of definiteness, i.e., definite articles (Mladenova 2007: 297–300). A decrease in frequency may be taken as an indication of the transition between the Balkan Slavic into the article-less non-Balkan South Slavic varieties, Serbian, and further BCMS varieties. However, little is known about their grammatical status in contemporary transitional varieties. The literature tends to provide brief and superficial descriptions, often using the analogy with the other Balkan Slavic languages (cf. Tomić 2006; Friedman 2006), or provide undetermined definitions, such as that of Ivić (1985: 116–17), describing them as articles with a strong demonstrative meaning. No sources provide sufficient details or empirical analysis

[1] Since these Slavic languages observe an SVO word order, one would expect prepositive rather than postpositive articles (Greenberg 1963). Word order has not been a part of this study.

The present paper presents an empirical analysis of their usage in the Timok variety of the Torlak zone, using the corpus of authentic spoken data from the region. Apart from the variation observed in the historical transitional varieties, Timok is presently affected by a strong influence from the dominant standard Serbian variety that is reflected in contemporary variation in the use of postpositive demonstratives (Vuković et al. 2023). All things considered, the goal of the present analysis is to look into different grammatical aspects of the distribution of these particles in order to reveal their grammatical status with respect to the evolution from demonstratives into definite articles. For the sake of the argument, since the status of these demonstrative particles in Timok is unknown, we refrain a priori from categorizing them as articles, which is their more settled status in the other two languages. In the following, we shall use the term "short demonstratives" (SDs) to denote shorter, enclitic postpositive forms of demonstratives.

2. Short Demonstratives in Timok

Short demonstratives (SDs) are one of the most salient features of the Timok dialect. They are derived from three demonstrative stems: the speaker proximal *-t*, (1a), hearer proximal *-v*, (1b), and distal *-n*, (1c). SDs inflect for gender, (1a–e), and for case, (2). In Timok we find SDs in nominative/unmarked forms and in accusative/oblique/marked forms in plural and singular, although not all the forms of the paradigms that can occur are equally distributed. Vuković et al. (2023) show that a noun carrying an SD is less likely to be inflected than a bare noun.

(1) a. čovek-**at**[2]
man.M.SG.NOM-DEM
'the/that man'

b. čovek-**av**
man.M.SG.NOM-DEM
'the/this man'[3]

c. čovek-**an**
man.M.SG.NOM-DEM
'the/that man yonder'

d. žena-**ta** (-va/-na)
woman.F.SG.NOM-DEM
'the/that woman (this/yonder)'

e. polje-**to** (-vo/-no)
field.N.SG.NOM-DEM
'the/that field (this/yonder)'

[2] Phonological variants exist.

[3] The translations provided here are used to keep with the practice in previous literature regarding the interpretation of the meaning and function of SDs and are not intended to bias the reader at this stage in the paper. As will be revealed later, based on the findings of this study, the *t*-form can indeed be translated as an article. Regarding the other two SD forms, while the *v*-form has occasional anaphoric uses, it would be more accurate to translate the *v*- and *n*-forms as demonstratives.

(2)	Traže	na	čoveka-**toga**	ličnu	kartu.[4]
	ask.3PL.PRES	on	man.M.SG.ACC-DEM.ACC	personal	card.F.SG.ACC

‘They are asking for that man’s ID.’

The distribution of SDs within the noun phrase resembles the Balkan Slavic pattern: they are postpositioned to their host and agree with it in gender, number, and case; see example (3).

(3)	Unuk-**at**	sadi	višnje-**te**.
	grandson.M.SG.NOM-DEM	plant.3SG.PRES	cherry.F.PL-DEM

‘The grandson is planting the cherries.’

In nominal expressions containing modifiers, SDs take the second position and attach to the left-most modifier of the noun, as in (4).

(4)	Moja-**na**	unuka	ima
	my.F.SG.NOM-DEM	granddaughter.F.SG.NOM	have.3SG.PRES
	mladu	babu.	
	young.F.SG.ACC	grandmother.F.SG.ACC	

‘My granddaughter has a young grandmother.’

The variation of SDs in Timok might be due to non-linguistic factors, owing to the fact that the Timok variety is influenced by standard Serbian, which does not use SDs. This variation has been partially examined by Vuković and Samardžić (2018), who have found that SDs are used more in remote areas, far from urban centers, where people have little contact with the standard language. Their use has also been related to other extralinguistic factors, such as gender and age, with women and older speakers tending to use SDs more frequently (Vuković et al. 2023).

The large variability observed in Timok implies that SDs are not an essential element of the noun phrase. This raises the question of whether their usage is completely unsystematic or whether there might be a pattern that goes beyond the explanation offered by geographic or social factors. The present analysis aims to investigate the possible existence of a systematic pattern in the linguistic domain by examining the distribution of SDs at the level of the noun phrase, as well as their semantic aspect and their use in the referential structure.

[4] The examples given throughout the paper are extracted from the Spoken Torlak dialect corpus 1.0 (http://hdl.handle.net/11356/1281; Vuković 2020; see also Vuković 2021 and Miličević et al. 2023) and belong to the Timok variety unless stated otherwise.

3. Analysis of the Usage Patterns of Short Demonstratives in Timok

In the absence of previous analyses of SDs in Timok, we may address this question by turning to the surrounding South Slavic varieties in which this phenomenon has received more ample treatment, or we could consider more general tendencies observed crosslinguistically. SDs have fully grammaticalized into definite articles in other Balkan Slavic languages (Bulgarian and Macedonian), originating from adnominal demonstrative pronouns (ADPs). Modern Bulgarian standard and most varieties know only one form of the SD. In Macedonian standard and dialects, on the other hand, there are three forms (not all of which function as articles, see §3.1; Topolinjska 2006). These reflect the three deictic forms of ADPs, as in Timok. Mladenova (2007) explains how the process of grammaticalization from an ADP to a definite article occurred in Bulgarian and Macedonian by analyzing pre-standardized Bulgarian texts. In this diachronic process, the first post-positioned occurrences of demonstratives were optional anaphoric markers, which then became more frequent and became obligatory markers of definiteness in word-final position.[5]

In what follows, various aspects of the use of SDs in Timok will be discussed. The distribution of different demonstrative forms and their referential use is analyzed in section 3.1. The distribution of SDs across different types of nouns is addressed in section 3.2, while section 3.3 deals with the position and function of the SD within the noun phrase. In order to investigate general tendencies of the use of SDs in Timok, semantic, noun-phrase-internal criteria, as well as discourse-related criteria, will be used and tested in the corpus as a whole. The choice of linguistic parameters in this paper was partially determined by the structure of the data used. Apart from their relevance for the research question, linguistic criteria were chosen such that they can be processed automatically or semi-automatically based on forms found in the text. The analysis of semantic components of definiteness, such as, for example, inclusiveness or uniqueness, would require detailed and complex manual assessment of the context of each example—a very time-consuming task that goes beyond the methodological scope of corpus linguistics.

[5] The grammaticalization process of definite articles in Bulgarian and Macedonian coincided with the loss of grammatical case, with strong indications of direct causality between the two grammatical processes (Mladenova 2007). Initially, SDs in Old Church Slavonic and early stages of Bulgarian were marked for case, but inflectional markings were lost over time (Mladenova 2007; Šimko 2020). However, this aspect will not be addressed in this article. For more on the interaction between case inflection and SDs in Timok, see Vuković et al. 2023.

The analysis was performed in the Spoken Timok dialect corpus[6] (Vuković 2020; see also Vuković 2021 and Miličević et al. 2023), based on transcripts of fieldwork interviews recorded with the local population in Timok between 2015 and 2018. The fieldwork was conducted within the project "Guardians of the Intangible Heritage of the Timok Vernaculars"[7], including a total of 12 researchers with backgrounds in linguistics, anthropology, ethnography, folklore, and literature. Field researchers conducted semi-structured interviews and focused on various aspects of immaterial culture, such as oral history, biographical narratives, and traditional culture. The collection methodology produced long stretches of natural speech, which allows for analysis of language use. Data was gathered from speakers in many different locations across the whole area, so as to enable the study of inter-speaker and areal variation. Audio and video materials and interview protocols are kept in the Digital Archive of the Institute for Balkan Studies in Belgrade. Selected edited videos can be viewed on the YouTube channel "Terenska Istraživanja"[8].

The Spoken Timok dialect corpus encompasses a total of about 500,000 tokens, 446,000 tokens of speech by 165 dialect speakers in 63 locations and 54,000 by researchers. Corpus compilation optimized analysis of the non-standard Timok vernacular and internal language variation by making it possible to select at least one representative speaker from evenly distributed locations across the region. The corpus is not internally demographically balanced. Although both genders are included, the majority of the speakers in the corpus are elderly women (101 speakers with around 370,000 tokens), as they are carriers of the most non-standard Timok variety and thus chosen as the focus of data collection. They were also indirectly targeted in the process of the linguistically motivated data sampling for the corpus, with the goal of representing non-standard dialectal features (as described in Belić 1905; Stanojević 1911; Bogdanović 1979; Dinić 2008: ix–xxiii). To create a more balanced sample and allow for analysis of variation across generations, a sample of high-school students was added to the corpus. While the observer's paradox is always a challenge, the researchers tried to minimize it by increasing the length of interviews, as well as by conducting interviews in the dialect and guiding participants towards more personally engaging topics, depending on their personal inclination.

The researchers used a semi-phonetic approach in order to transcribe non-standard language features. The corpus contains automatic part-of-

[6] The official name is the "Spoken Torlak dialect corpus 1.0" (https://www.clarin.si/repository/xmlui/handle/11356/1281).

[7] "Čuvari nematerijalne batine timočkih govora", financed by the Ministry of Culture and Information of the Republic of Serbia.

[8] Available on YouTube at https://www.youtube.com/channel/UC4EpCSAnEb2RIsIRY7pfNdQ. Last accessed 3 August 2022.

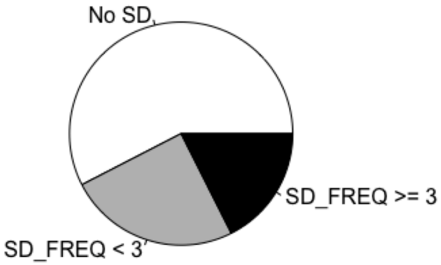


Figure 1. The distribution of SD frequency across speakers (per 1,000 tokens)

speech annotation and lemmatization performed using a custom model of the ReLDI tagger that was based on a manually annotated sample of 27,000 tokens (Vuković 2019; Ljubešić et al. 2016) (for more details regarding the corpus creation, see Vuković 2021).

Tags for words hosting an SD were manually verified in the corpus and used as such in the analysis. For the analysis, 1,313 examples of SDs uttered by dialect speakers were extracted (researchers' production was excluded). As mentioned earlier, there is a great deal of variation in the corpus when it comes to the use of SDs. To illustrate this, out of 165 speakers, only 70 speakers used SDs, and 39 speakers used 3 or more SDs per 1,000 tokens, as shown in Figure 1.[9] None of those speakers were in the group of high-school students. As mentioned above, previous research has shown that SDs are used much less by men and younger speakers (Vuković et al. 2023).

3.1. Demonstrative Stem and Type of Reference

Timok SDs have a tripartite reference differentiation, just like demonstrative pronouns: the speaker-proximal *v*-form, from the demonstrative *ovaj* 'this', the hearer-proximal *t*-form, from the demonstrative *taj* 'that' (sometimes described as distal), and the distal *n*-form, from the demonstrative *onaj* 'that over there, yonder', which signifies referents far from both the speaker and the hearer. In Timok all three demonstrative pronouns are used postpositively as short demonstratives, as shown in (5).

[9] Bear in mind that the use of SDs was one of the criteria in the selection process when creating the corpus sample, being one of the distinguishing dialectal features. Those who use SDs were strongly favored. On the one hand, it can be assumed that the proportion of speakers who use SDs within the entire population of Timok would be smaller. On the other hand, it is difficult to judge to what extent the observer's paradox affects the use of SDs, given their salience, and it could be the case that more people actually use them when researchers are not present.

(5)	a.	**taj**	čovek	/	čovek-**at**
		that.M.SG.NOM	man.M.SG.NOM		man.M.SG.NOM-DEM
		‘that man’			‘that/the man’
	b.	**ovaj**	čovek	/	čovek-**av**
		this.M.SG.NOM	man.M.SG.NOM		man.M.SG.NOM-DEM
		‘this man’			‘this man’
	c.	**onaj**	čovek	/	čovek-**an**
		that.M.SG.NOM	man.M.SG.NOM		man.M.SG.NOM-DEM
		‘that man yonder’			‘that man yonder’

Belić (1905: 443–44) states that in Timok the *t*-stem is used with a definite and demonstrative meaning, while the other two, *v*- and *n*-stem, have only demonstrative meaning and are less often postponed. He provides no examples of this distinction, nor empirical foundations, but his claim offers two premises: (i) *t*-stem is the one most frequently used as an SD, and (ii) there is a difference between demonstrative and definite meaning related to different forms of SDs. The first premise is in accordance with the other two Balkan Slavic languages which have fully grammaticalized definite articles.[10] The *t*-stem is the only root for the definite article in Bulgarian (Mladenova 2007: 94). In Macedonian the *t*-stem is used as an article, but the other two are not (Koneski 1967: 228–32; Topolinjska 2006; Karapejovski 2020: 168–80; Boronnikova 2014, cf. Friedman 2001). If Timok should indeed display the same tendency as Bulgarian and Macedonian, we could expect that the *t*-stem short demonstrative would be used more frequently than the other two in comparison to the frequency of the ADP. To test this, normalized frequencies of each form of the SD will be compared with the normalized frequency of ADPs (normalized per 10,000 nouns) and the statistical difference between them using a chi-square test.

Regarding the second premise, the shift from demonstratives to the definite article is indicated by the increase in the anaphoric use of demonstratives or demonstrative-like elements (Greenberg 1978; Diessel 1999). This is found to be true in languages across the world (Greenberg 1978; Diessel 1999), and more importantly, it has been confirmed in the earlier stages of Bulgarian (and generalized to other Balkan Slavic languages) where anaphoric use of demonstratives gave rise to the definite article (Mladenova 2007). In the case of Macedonian, a language with a tripartite deictic reference expressed in both ADPs and SDs, like in Timok, the *t*-form is used as a definiteness marker, while the other two preserve a demonstrative meaning (Koneski 1967: 228–32; Topolin-

[10] For expression of definiteness in Old Church Slavonic, including SDs, see Karamfilova 1998.

jska 2006, cf. Karapejovski 2020: 168–80; Boronnikova 2014). The distinction in Macedonian is made between a deictic meaning, linked to demonstratives, and anaphoric meaning, linked to articles. Thus, *v*- and *n*-forms are deictic elements, equal to ADPs, while the *t*-form is said to perform an anaphoric function and can therefore be classified as an article (Topolinjska 2006; Karapejovski 2020: 168–80; Boronnikova 2014). A similar distinction is found in more general literature. That is, demonstratives need to match the referent to a perceptible object; the definite article loses this matching constraint and can rely on general knowledge and the discourse (Hawkins 1978: 149–58).

Furthermore, as grammaticalization advances towards marking definiteness in Bulgarian and Macedonian, generic nouns can bear an article (Mladenova 2007: 93). Also, articles can be used in nominalizations (Tomić 2006: 58, 90).

With the goal of empirically analyzing the referential function that short demonstratives perform in Timok, they will be manually categorized according to the type of reference: deictic, which corresponds to demonstratives, and anaphoric, corresponding to articles. Deictic referencing relates to spatial deixis, evident directly or from the content of the surrounding narrative (Diessel 1999: 35–46; Levinson 1983: 61–96), as well as from metaphorical expression of deixis, such as emotional distance (Lakoff 1974). Anaphoric reference points to referents already mentioned in the discourse or known to exist based on speakers' shared knowledge. Another layer of analysis relates to the distinction between generic versus non-generic interpretation of nominals. This categorization will be combined with the demonstrative stems in order to determine which form of SD is used anaphorically and which deictically.

3.1.1. Analysis

For the analysis of the frequency of use of demonstrative stems in SDs and ADPs, each occurrence of SDs and ADPs was extracted from the corpus and marked with a respective value. The occurrences of SDs were retrieved using the manually verified PoS tags (see §3). ADPs were extracted and marked automatically using PoS tags and word forms. In order to compare the use of demonstrative stems across the whole corpus, the absolute frequencies of SDs and ADPs were segmented based on the type of demonstrative stem (*-t*, *-v*, *-n*) and normalized per 10,000 nouns. A chi-square test was used to compare frequency distributions between ADP and SD forms to determine whether there are differences in how each of the demonstrative stems is used depending on how they appear with the noun.

When it comes to the type of reference of words containing an SD, the data was annotated manually for deictic or anaphoric reference and generic or non-generic. Regarding the former, some referents are both deictic and anaphoric, as they can be identified in the physical space but also involve referents that have been prominent in the previous discourse. Annotation was based on text alone; video materials were not found necessary for the analysis. Pearson's chi-square test was used to determine whether there is a difference in frequencies departing from a uniform distribution among variables. In assessing the variation of the use of different demonstrative stems for deictic or anaphoric purposes—i.e., in the analysis of interdependence between the use of demonstrative stems and types of reference—the method of linear regression was used. This measure serves to indicate the intensity of association, or whether the value of one variable can be predicted based on the value of the other variable. The dependent variable was the demonstrative stem, differentiating between the *t*-stem and the other two stems: *t*-stem being one value, *v*- and *n*-stem another. The independent variable was the type of referential usage—deictic or anaphoric. In this case, two linear regression analyses were performed: one to estimate the relationship between the *t*-stem and anaphoric reference and another one for *v*- and *n*-stem jointly and deictic reference.[11]

3.1.2. Results

Among the three SD forms, the *t*-stem is used most frequently, as evidenced by normalized frequencies across the whole corpus (see Table 1).

Table 1. Frequencies of demonstrative stems used as ADP and SD normalized per 10,000 nouns

	***t*-stem**	***v*-stem**	***n*-stem**
ADP	146.29	24.60	146.69
SD	146.56	75.53	4.23

The variation between the use of different stems as an SD or ADP, assessed with a chi-square test, showed a significant result (*x*-squared = 104.7, df = 1, *p*-value < 0.001). From the frequencies, we see that the *v*-stem is used more frequently as an SD than as an ADP, while the *n*-stem is used very rarely as an SD, compared to the equivalent ADP and compared to other SD forms.

When it comes to the type of reference of different forms of SDs, the data from the corpus as a whole shows that the *t*-stem is used mainly for anaphoric reference, while the *v*-stem and *n*-stem are mainly used deictically. At the same time, there are some mixed cases that offer both a deictic and

[11] For chi-square test, 'chisq.test()' function was used, while for linear regression, function 'lm()' was used from the R package Stats (R Core Team 2022).

an anaphoric interpretation. In example (6), the referents marked with an SD refer to referents previously mentioned in the discourse, while also bearing a reference to an object easily identifiable in the physical space.

(6)	Ima	reka	pa	se	pravi
	have.3SG.PRES	river.F.SG.NOM	SO	REFL.ACC	make.3SG.PRES

vada. […]	Ima	gore	vrelo […]	dole
canal.F.SG.NOM	have.3SG.PRES	up.there	spring.N.SG.NOM	down.there

u	reku-**tu**
in	river.F.SG.ACC.DEM

'There is a river up there, so a canal is made. […] There is a spring up there […] down by the river'

Raw frequencies of the SD form classified according to the stem and type of reference are shown in Table 2.

Table 2. Demonstrative stems and the type of reference (raw frequencies)

	Only D	Only A	D and A	Total
t-stem	15	1000	90	1105
v-stem	154	8	5	167
n-stem	29	0	3	32

The use of the *t*-stem is strongly preferred with the anaphoric type of reference across speakers, as indicated by linear regression (F-statistic = 4.466e+04 on 1, df = 70, *p*-value < 0.001). The use of *v*- and *n*-stems was strongly favored for deictic types of reference (F-statistic = 792.7 on 1, df = 70, *p*-value < 0.001).

Out of 72 speakers who use SDs in the whole corpus, 19 speakers used the *n*-form, 38 speakers used the *v*-form, and 67 speakers used the *t*-form of the SD (meaning that some speakers did not use the *t*-form, but the other two forms instead). Moreover, rarely do speakers use all three forms; only one speaker (TIM_SPK_0028) uses all three forms frequently ($N_{t\text{-form}} = 30$, $N_{v\text{-form}} = 54$, $N_{n\text{-form}} = 10$). The majority of speakers use the *t*-form dominantly or exclusively, especially those who make frequent use of SDs.

The relationship between the two variables was explored further using linear regression, and it was found that, interestingly, speakers who use the typically deictic SDs tend to use SDs deictically overall, including the *t*-stem.[12]

[12] These findings are the result of an analysis across speakers, where the independent variable was the total number of *v*- and *n*-stems, and the dependent variable was

This also indicates that others exhibit a tendency towards a more general anaphoric use, using only the *t*-form with strong anaphoric preference. This suggests that some speakers have a more demonstrative-like use of SDs, while others have a more article-like use of SDs.

Looking into particular cases of individual speakers might reveal something about the mechanisms of grammaticalization. As an illustration of individual cases, the speaker TIM_SPK_0002, who uses all three forms, but the *t*-form dominantly ($N_{t\text{-form}}$ = 41, $N_{v\text{-form}}$ = 6, $N_{n\text{-form}}$ = 2), tends to use SDs anaphorically (41 anaphoric uses out of 50). Another speaker, TIM_SPK_0005, uses 38 SDs, 37 of which are the *t*-form, all used anaphorically; speaker TIM_SPK_0011 uses 78 SDs, 77 of them are *t*-form, 76 of which are used anaphorically; speaker TIM_SPK_0011 uses 90 SDs, all *t*-forms used anaphorically. This trend is repeated with other speakers (e.g., TIM_SPK_0035, TIM_SPK_0040, TIM_SPK_0061). By contrast, the speaker TIM_SPK_0028 mentioned above uses *v*- and *n*-forms deictically but also shows 7 occurrences of deictic *t*-form. The correlation between the use of the *v*- and *n*-form and the deictic use of SDs, including the *t*-form, is more striking with the speakers who use SDs less frequently. Some speakers who use SDs less frequently often use them deictically. For instance, speaker TIM_SPK_0046, who uses 10 SDs in total ($N_{t\text{-form}}$ = 9, $N_{n\text{-form}}$ = 1), shows 8 deictic uses; speaker TIM_SPK_0094, a total of 13 SDs, all *t*-form, out of which 10 are used deictically; speaker TIM_SPK_0132, who uses 4 SDs ($N_{v\text{-form}}$ = 3, $N_{n\text{-form}}$ = 1), uses them only deictically. As shown in the above correlation, when a speaker uses the *t*-form dominantly, they also use SDs anaphorically. Moreover, the data suggests that, once the *t*-form becomes more frequent, anaphoric usage takes over and the other two forms decrease in frequency. More importantly, this shift happens in individual speakers, which suggests that grammaticalization occurs in individual speakers or individual grammars.

Regarding genericity, all instances of SDs in the corpus are non-generic, which means that SDs in Timok are used for anaphoric or deictic marking only. Even when used with mass or collective nouns, they have either been explicitly elicited by the previous discourse or clearly identifiable within the discourse or shared knowledge. There are no truly generic usages of SDs observed in the corpus.

3.2. Type of Noun

In Macedonian and Bulgarian, SDs occur with a variety of noun classes, including count, mass, and generic nouns (Mladenova 2007: 4; Tomić 2006: 58–59, 90–91), each representing a different selection scope, being able to attach to

whether the *t*-stem was used anaphorically (F-statistic = 7.164, $df_N = 1$, $df_D = 70$, *p*-value < 0.01).

nouns denoting singular units, multiple units, mass, or a genus. They pertain to different categories regarding criteria such as uniqueness, identifiability, inclusivity, genericity, and so on, depending on how they refer to real-world concepts (see Lyons 1999: 7–15). When it comes to the pragmatic and semantic notion of definiteness, Mladenova (2007: 4–5) singles out identifiability as a linguistic universal (based on Lyons 1999: 278–318), whereas some languages may further develop meanings such as inclusiveness, genericity, specificity, etc. The cycle involves the expansion from identifiability (pertaining to demonstratives) to inclusiveness (pertaining to articles), and further to genericity. As Mladenova notes, the Bulgarian and Macedonian *t*-article has evolved into a genericity marker.

The occasional use of SDs in Timok may imply that not every noun can bear one, that certain types of nouns appear more frequently than others, and that there may exist restrictions in the lexical domain. The focus of this section is to examine whether the grammatical or lexical criteria of nouns can indicate their likelihood of hosting an SD in Timok relative to their meaning. This further relates to their status in the transition between demonstratives and articles.

As has already been described in the previous section, in Timok there are no true generics used with an SD, thus the transition may fall between the notions of identifiability and inclusiveness. In terms of nominal classification based on lexical semantics, this transition can be observed in the distinction between count and mass nouns as well as concrete and abstract nouns. Within the two distinctions, count and concrete nouns are more easily identifiable because of their quantifiable and material properties and thus reflect a demonstrative-like meaning. On the other hand, the immaterial nature of abstract nouns makes them less easy to identify conceptually, while mass nouns elicit the inclusiveness criterion, given that they do not refer to singular entities. These two distinctions are therefore taken as representative for situating the SD in Timok on the grammaticalization path between demonstrative and article. The analysis focuses broadly on the chances for a noun to occur with an SD and, more specifically, on whether there is a significant difference in frequency between count and mass nouns and concrete and abstract nouns.

3.2.1. Analysis

In order to determine the probability of each noun occurring bare or with an SD, the confidence interval was measured for the occurrence of lemmas for bare nouns and nouns hosting SDs in the corpus.[13] All noun lemmas in the corpus were examined and categorized into bare nouns and nouns with SDs, and the relative proportion of each lemma in both categories was calculated.

[13] R package CI was used (Fneish 2021).

For the analysis of the semantic criteria of count vs. mass and concrete vs. abstract nouns, each lemma was labeled manually. Only common nouns were included. Since the list of all noun lemmas in the corpus is large (14,420 lemmas), a smaller number of frequent lemmas were selected for analysis: all lemmas hosting an SD and bare nominal lemmas that occur at least 10 times in the corpus. The subset had a total of 1,278 lemmas, out of which 162 were proper nouns, resulting in a sample size of 1,116 lemmas. The data was then analyzed using linear regression,[14] measuring the relationship between the frequency of nouns hosting an SD and the variables representing countable (1 = yes, 0 = no) and concrete (1 = yes, 0 = no).

3.2.2. Results

The total number of noun lemmas occurring bare is 14,420, while the total number of lemmas occurring with an SD is 410. Relative proportions in each category reveal a notable difference: the confidence interval for the likelihood of occurrence of bare noun lemmas ranges between 97.5% and 97.9% (95% CI), while for nouns bearing SDs, the range is between 2.07% and 2.52% (95% CI), which means that a lemma is much less likely to occur carrying an SD. The quantitative differences between the two categories are illustrated in Table 3.

Table 3. Descriptive statistics and confidence interval for lemmas in each category

	Max (abs freq)	**Mean (abs freq)**	**SD (abs freq)**	**CI LL**	**CI UL**
Bare noun	1,400	5.48	33.35	**97.50%**	**97.90%**
Noun + SD	27	0.07	0.77	**2.07%**	**2.52%**

The frequency rank distribution among the two categories is not equal. The most frequent lemmas in each category and their frequencies are shown in ~~Table 4~~. Figure 2

[14] Function 'glm()' was used from the R package Stats (R Core Team 2022).

Table 4. Linear regression statistics

	B (SE)	Odds ratio	*t*-value	*p*-value
Count	0.57 (0.21)	1.77	2.66	<0.001
Concrete	1.04 (0.23)	2.83	4.45	<0.001

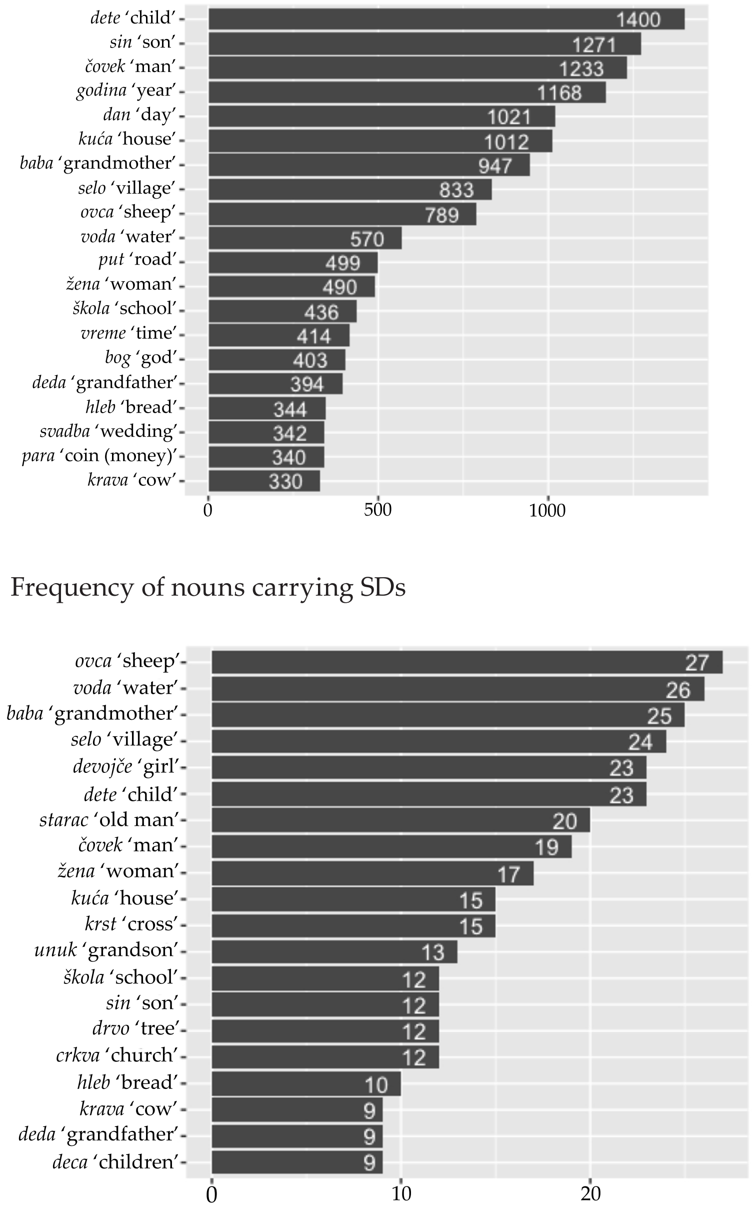
Notice the actual nouns displayed on the *y*-axes and how the lexical scope and the order do not correlate. For instance, the maximum absolute frequency for a bare noun is 1,400, observed with the noun *dete* 'child' (ranked 6th in the marked category), while the maximum absolute frequency for a noun hosting an SD is 27, observed with the noun *ovca* 'sheep' (ranked 9th in the bare category). The ranking discrepancy is found to reflect the differences in semantic selection criteria that are illustrated in Figure 2 on the following page.

When it comes to the analysis of the semantic criteria, both distinctions were revealed to be statistically significant, according to linear regression. The odds are 1.77 higher for SDs to occur with count nouns than with mass nouns, and 4.45 higher for SDs to occur with concrete nouns than with abstract nouns (see Table 4). These findings provide further support for a similar conclusion in the previous section. The SD in Timok is not at the same grammatical level as in Bulgarian and Macedonian. The fact that it tends to co-occur with concrete and count nouns pertains more to its deictic roots than to the abstract notion of definiteness.

3.3. Distribution in the Noun Phrase

There is a clear initial difference in the structure of the noun phrase, especially when it comes to the class of determiners, between Serbian, located on the western border of the Torlak region, and Bulgarian and Macedonian, located on its eastern border. In standard Bulgarian and Macedonian, articles, in the form of SDs, are an obligatory element of nominal expressions with a definite, i.e., identifiable, interpretation (except inherently definite nouns such as proper names, toponyms, etc., although they can be marked as well; Tomić 2006). On the other hand, in standard Serbian and surrounding Serbian varieties, definiteness is not grammatically marked as in Bulgarian, and determiners are not an obligatory element of the noun phrase (Stanković 2017). Given the lower frequency of SDs in Timok, their usage can be expected to reflect earlier stages of the grammaticalization process observed diachronically in Bulgarian and Macedonian. Apart from the analogy in frequency, distributional patterns within the structure of the noun phrase can be used to assess their grammatical status. Their linear position and co-occurrence with other

Figure 2. Lemma frequency distribution for bare nouns and nouns carrying an SD (absolute frequency)



a. Frequency of bare nouns

b. Frequency of nouns carrying SDs

nominal elements can locate SDs in the hierarchy of nominal constituents and indicate their meaning and functional properties.

In Bulgarian and Macedonian, the SD pertains to the functional layer of the NP. It exhibits minimal selection restrictions for its host, as demonstrated by Dimitrova-Vulchanova and Vulchanov 2010 (cf. Zwicky 1977; Zwicky and Pullum 1983). This means that it can be hosted by different constituents within a nominal expression: adjectival modifiers such as possessive pronouns and some numerals (Topolinjska 2009), quantifiers (e.g., *many* and *all*), and the head noun (Dimitrova-Vulchanova and Vulchanov 2010). The selection restrictiveness (or lack thereof) is found to correlate to the definiteness status of the SD. The less restrictive it is in the selection of its host, the less it has the immediate deictic meaning of the ADP, and the more it has the meaning of inferred identifiability of the article (Dimitrova-Vulchanova and Vulchanov 2010). In the hierarchy of nominal modifiers, those positioned to the left are ranked higher within the NP, with quantifiers being the leftmost and highest-ranked. Elements in the leftmost periphery of the NP are the last to be eligible as hosts for an SD in the grammaticalization process. This progression towards the left indicates a shift in grammatical function: ADP > SD attaching to nouns > SD attaching to adjectival modifiers > SD attaching to high-ranking modifiers such as quantifiers. Consequently, the attachment of an SD to the leftmost elements of the nominal expression signals its evolution from a deictic ADP to a marker of definiteness.

The variation in the use of the SD in Timok may suggest that it has not fully grammaticalized into a definiteness marker and that, syntactically speaking, it remains in the grammaticalization phase of the anaphoric article or even the deictic element. Current research on Timok has revealed that SDs appear with nouns without modifiers more frequently and that they attach more frequently to nouns than to other parts of speech (Vuković et al. 2023).

The distribution of the SD within the NP, and more precisely, its phrase-internal selection pattern, is used to analyze the status of the SD with respect to its development from a demonstrative into a definite article. Should it attach to high quantifiers such as *many* and *all*, it can be interpreted as a definite marker belonging formally to the functional layer of the NP. More restrictive host selection is taken as an indication of its lower grammatical status.

3.3.1. Analysis

We searched for nominal expressions containing left modifiers (adjectives, possessive pronouns, demonstrative pronouns, numerals, and quantifiers). The extracted examples were first classified according to whether the NP contained an SD. Those that did were then analyzed for the particular left constituents they contained and which one of them was hosting the SD. Examples of nominal expressions were extracted from the corpus using PoS tags. Examples of occurrences of SDs were extracted from the corpus using manually verified PoS tags (see §3). These were further manually processed to search and account for the occurrence of SDs with different constituents of the nominal expression. This part of the study did not allow for statistical analysis, owing to the small sample size.

3.3.2. Results

In the Timok sample, SDs occur rarely in quantified nominal expressions (N_{quant} = 9), and only with numerals. In the one occurrence of a cardinal numeral as a quantifier, the SD is on the noun, (7a). The adjectival use of numerals is more frequent (N_{ordnum} = 5), and in that case, the SD attaches to the numeral functioning as an adjectival modifier, (7b). There are four occurrences of quantifiers like *oba/obojica* 'both'. In two instances, the quantifier hosts the SD, as in (7c), while in the other two, the SD is attached to the quantified noun, as in (7d). In general, SDs tend to occur with lower numerals, which exhibit adjectival syntax. Universal quantifiers, such as *many* and *all*, do not occur with an SD.

(7) a. tri ovce-te
three sheep.F.PL.NOM-DEM
'three sheep'

b. druga-ta noga
other.F.SG.NOM-DEM leg.F.SG.NOM
'the other leg'

c. obojica-ta sina
both.F.SG.NOM-DEM son.M.SG.GEN
'both sons'

d. oba starca-voga
both old.man.M.SG.GEN-DEM.M.SG.GEN
'both old men'

In examples with an adjectival modifier to the left of the noun in the initial position within the nominal expression (N_{adj} = 13), the SD appears on the adjective, as in (8a). In instances of double determination with the structure ADP + ADJ + N attested in the corpus (N = 2), the SD is again hosted by the adjective, as illustrated in (8b).

(8) a.

stara-ta	žena
old.F.SG.NOM-DEM	woman.F.SG.NOM

‘the old woman’

b.

toj	srednji-ti	dan
that.M.SG.NOM	middle.M.SG.NOM-DEM	day.M.SG.NOM

‘that middle day’

In 27 phrases with a possessive pronoun in the initial position, 26 show an SD on the possessive. The one instance where this is not the case has a structure that includes an adjective to which the SD attaches: POSS + ADJ + SD + N. Among the possessives, three examples exhibit an SD on both the noun and the possessive, while one hosts an SD only on the possessive but not the noun.

Out of 52 instances of double determination involving a demonstrative and an SD, demonstrative stems coincide 30 times, while in 12 examples, they are different. Out of those 12 examples, 10 involve a *t*-stem SD (19 out of the 52 include an *n*-stem demonstrative).

Upon examining the examples, it turned out that not all modifiers in the corpus bear an SD. Quantifiers such as *many* and *all* seldom co-occur with a noun or another element hosting an SD, but they themselves never host an SD (in such phrases, the noun is the host). Demonstratives co-occur with SDs but never host them. The sample suggests that in Timok only adjectival modifiers can bear an SD. Coming back to what we know from Bulgarian and Macedonian, this implies that SDs in Timok do not have the status of definite articles, but rather an anaphoric function, as they are not hosted by universal modifiers and select only adjectival elements as hosts. The insight based on double determination phenomena suggests that the *t*-stem carries the anaphoric meaning more than the other two, with the *n*-form being the most deictic one, confirming the findings on the type of reference from §3.1.

4. Discussion

The genesis of the definite article in Balkan Slavic languages follows a cross-linguistic observation that the ADP is a common root for the grammaticalization of articles. As Greenberg (1978: 61) finds, ADPs, being markers with purely deictic reference, are grammaticalized into markers with anaphoric discourse reference and are then extended to markers of definite elements. The transition from an ADP is initially marked by the increased anaphoric use of demonstratives (or demonstrative-like particles) (see Diessel 1999; Heine and Kuteva 2006: 110). The variation found in Timok, and the non-obligatory nature of the SD that it includes, fits into what Lyons (1999: 52) describes as "optional" usage of article-like demonstratives that is found in some languages where article-like elements occur only occasionally.

Observations from a broader Slavic perspective (Mendoza 2014) show that the expansion of article-like usage of demonstratives is propelled by the increasing need to mark an anaphoric NP in order to connect it with its antecedent or an exophoric context. The usage of these particles differs between the Slavic languages described by Mendoza (2014): Polish, Czech, Upper Sorbian, and 17th-century Russian texts written by Avvakum. However, as in Timok, they all display a certain degree of optionality depending on the context. Following the criteria applied by Mendoza (2014), the SD in Timok seems to show indications that the article is currently in an anaphoric grammaticalization stage, given that it is used with possessive NPs and can occur with proper nouns.

This is further in line with the findings presented here. That is, although "optional", the use of SDs in Timok reveals a pattern that points to a set of characteristics indicating a specific phase in the grammaticalization process, namely that of an anaphoric article. SDs in Timok do not show clear indications for the status of a full-fledged definite article, as is found in Bulgarian and Macedonian. It has been substantiated by findings that SDs tend towards concrete and countable nouns, an indication that they maintain some demonstrative semantic elements. Within the NP, they do not take the typical position of the definite article, as they do not co-occur with other determiners, such as quantifiers, in contrast to the NP structure in Bulgarian and Macedonian.

As the increase in the frequency of the SD may be taken as an indicator of its advancement towards proper article status, the data presented here allows us to speculate that certain speakers in Timok are located further on that path than others and that this may altogether serve as an argument for a general tendency in the Timok variety.

We can speculate that the high variability in the use of SDs in recent years is affected by the decreasing number of speakers of the highly non-standard Timok variety. The decrease in speakers is particularly due to the depopulation of remote rural areas and migration to urban areas, where the standard is more prevalent. This assumption is indirectly indicated by the lesser use of several dialectal features by younger speakers (Vuković et al. 2023), given that the younger population is centered around cities and key infrastructure. Another factor linked to the age effect is that several salient dialectal features show a high degree of mutual correlation in terms of variation across the population (Vuković et al. 2022). However, the specific changes in the Timok population size and the influence of these changes on language have not been studied.

The data analyzed provides insight only into the synchronic situation in Timok and does not allow for a diachronic perspective. Furthermore, the sample used here is not balanced, in that it includes mostly older speakers, the majority of whom are women. Despite clear indication that this is exactly the part of the population in Timok that uses SDs (Vuković et al. 2023), a more balanced sample could reveal tendencies across the younger population, including male speakers. A more balanced corpus could also allow for the consideration of other factors, such as education, mobility, etc. Finally, corpora provide insight into language use that is evidenced in a given sample, but not all possible natural language utterances are available, a limitation that can be minimized, but not eliminated, by sampling techniques.

5. Summary and Conclusion

The present study addresses the question of the status of short demonstratives in Timok in the process of grammaticalization from a demonstrative into a definite article. It uses insights from neighboring Bulgarian and Macedonian varieties, where this process of grammatical change has resulted in a fully grammaticalized definite article, as well as cross-linguistic insights into the process. In a sense, the analyses presented here elaborate on the rather vague description put forward by Pavle Ivić (1985: 116–17), stating that SDs in Timok are "used like articles with a strong demonstrative meaning".

This study was performed through an array of quantitative analyses, using a dataset compiled from interviews with contemporary speakers of the Timok variety. It uses pragmatic, semantic, and syntactic criteria and analyzes whether SDs are used anaphorically or deictically and how they are distributed in the noun phrase and sentence. The results show that although there is variation in the anaphoric and deictic use of SDs, the *t*-form of the SD is predominantly used for anaphoric referencing, while *v*- and *n*-forms are more commonly used deictically. The results also show that some speakers tend to use SDs more deictically than others. The analysis of semantic parameters such as countability vs. uncountability and concreteness vs. abstractness reveals that SDs prefer countable and concrete nouns, which is a counterindication for their definite status. Furthermore, the analysis of NPs hosting SDs shows that within a nominal expression, the SD attaches almost exclusively to adjectival modifiers, which suggests that it does not have the status of a functional element marking definiteness.

Considered within the context of the grammaticalization of demonstratives into definite articles that has occurred in Bulgarian and Macedonian, the results of this study indicate that short demonstratives in Timok have not reached the grammaticalization stage of the definite article. The increased use of the *t*-stem, as well as the common anaphoric use of the same morpheme, however, indicates that the process of grammaticalization is likely occurring (that SDs are not identical to adnominal demonstrative pronouns). Still, no indications have been found that this process has advanced beyond anaphoric usage. The same can be confirmed by other analyses regarding the type of noun selection and distribution within the NP.[15]

Sources

Vuković, Teodora. (2020) "Spoken Torlak dialect corpus 1.0 (transcription)". Slovenian language resource repository CLARIN.SI. Available at: http://hdl.handle.net/11356/1281. Last accessed 3 August 2022.

References

Belić, Aleksandar. (1905) *Dijalekti istočne i južne Srbije* [The dialects of eastern and southern Serbia]. Belgrade: Srpska Kraljevska Akademija.

Bogdanović, Nedeljko. (1979) *Govori Bučuma i Belog Potoka* [Dialects of Bučum and Beli Potok]. Belgrade: Institut za srpskohrvatski jezik.

Boronnikova, Natalija Vladimirovna. (2014) "Status trojnogo člena v makedonskom jazike" [The status of the tripartite article in Macedonian language]. *Filologičeskie nauki: Voprosy teorii i praktiki* 10(40): 60–65.

Diessel, Holger. (1999) *Demonstratives: Form, function, and grammaticalization*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Dimitrova-Vulchanova, Mila and Olga Mišeska Tomić. (2009) "The structure of the Bulgarian and Macedonian nominal expression: Introduction". Mila Dimitrova-Vulchanova and Olga Mišeska Tomić, eds. *Investigations in the Bulgarian and Macedonian nominal expression*. Trondheim: Tapir Akademisk Forlag, 1–23.

Dimitrova-Vulchanova, Mila and Valentin Vulchanov. (2010) "An article on the rise: Contact-induced change and the rise and fall of N-to-D movement". Anne Breitbarth, Christopher Lucas, Sheila Watts, and David Willis, eds. *Continuity and change in grammar.* Amsterdam: John Benjamins Publishing Company, 335–54. [Linguistik Aktuell/Linguistics Today, 159.]

———. (2011) "An article evolving: The case of Old Bulgarian". Dianne Jonas, John Whitman, and Andrew Garrett, eds. *Grammatical change: Origins, nature, outcomes*. New York: Oxford University Press, 160–78. DOI 10.1093/acprof:oso/9780199582624.003.0008.

Dinić, Jaksa. (2008) *Timočki dijalekatski recnik* [Dictionary of the Timok dialect]. Belgrade: Institut za srpski jezik SANU.

Fneish, Firas. (2021) CI Package (Confidence Interval), Version: 0.0.0.9000. Available at: https://github.com/firasfneish/CI-package.

[15] At the time of the publication of this paper, the author is affiliated with the Digital Society Initiative and Department of Computational Linguistics at the University of Zurich. Most of the work on this paper, however, was done during the author's tenure at the Slavisches Seminar, University of Zurich.

Friedman, Victor A. (2001) *Macedonian*. Durham, NC: SEELRC, Duke University. [SEELRC Reference Grammars.] Available at: http://www.seelrc.org:8080/grammar/pdf/compgrammar_macedonian.pdf. Last accessed 21 June 2021.

———. (2006) "Balkans as a linguistic area". Keith Brown, ed. *Encyclopedia of language and linguistics*. 2nd ed. Vol. 1. Oxford: Elsevier, 657–72.

Greenberg, Joseph H. (1963) "Some universals of grammar, with particular reference to the order of meaningful elements". Joseph H. Greenberg, ed. *Universals of language*. Cambridge, MA: MIT Press, 40–70.

———. (1978) "How does a language acquire gender markers?". Joseph H. Greenberg, ed. *Universals of human language 3: Word structure*. Stanford, CA: Stanford University Press, 49–81.

Hawkins, John. (1978) *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. 1st ed. London: Routledge. DOI 10.4324/9781315687919.

Heine, Bernd and Tania Kuteva. (2006) "The rise of articles". Bernd Heine and Tania Kuteva, eds. *The changing languages of Europe*. Oxford: Oxford University Press. DOI 10.1093/acprof:oso/9780199297337.003.0003.

Ivić, Pavle. (1985) *Dijalektologija srpskohrvatskog jezika: Uvod i štokavsko narečje* [Dialectology of the Serbo-Croatian language: Introduction and Shtokavian dialects]. Novi Sad: Matica srpska.

Joseph, Brian. (1992) "The Balkan languages". William Bright, ed. *International encyclopedia of linguistics*. Vol. 4. Oxford: Oxford University Press, 153–55.

Karamfilova, Petya. (1998) "Sredstva za izrazjavane na opredlenost v starija bulgarski knižoven ezik do XV–XVI vek" [Means for expressing definiteness in Old Bulgarian literary language in the 15th to 16th century]. Cenka Ivanova, Tošana Stojanova, and Ivan Xaralampiev, eds. *Bălgaristični proučvaniija* [Bulgarian studies]. Vol. 3. *Aktualni problemi na bugaristikata i slavistikata* [Current problems of Bulgarian and Slavic studies]. Veliko Tŭrnovo: Universitetsko izdatelstvo "Sv. Sv. Kiril i Metodij", 169–86.

Karapejovski, Boban. (2020) *Eksponentite na kategorijata obredelenost vo makedonskiot jazik* [Exponents of the definiteness category in the Macedonian language]. Ph.D. dissertation, Saints Cyril and Methodius University.

Koneski, Blaze. (1967) *Gramatika na makedonskiot literaturen jazik* [Grammar of the Macedonian literary language]. Skopje: Kultura.

Lakoff, Robin. (1974) "Remarks on 'this' and 'that'". *Proceedings of the Chicago Linguistic Society* 10: 345–56.

Levinson, Stephen C. (1983) *Pragmatics*. Cambridge: Cambridge University Press.

Lindstedt, Jouko. (2000) "Linguistic Balkanization: Contact-induced change by mutual reinforcement". Dicky Gilbers, John Nerbonne, and Jos Schaeken, eds. *Languages in contact*. Amsterdam: Rodopi, 231–46. [Studies in Slavic and General Linguistics, 28.]

Ljubešić, Nikola, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. (2016) "New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian". Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, eds. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: ELRA, 4264–70. Available at: https://aclanthology.org/L16-1676.pdf.

Mendoza, Imke. (2014) "Das Pronomen *tъ und seine Rolle bei der Grammatikalisierung von Definitheit im Slavischen" [The pronoun *tъ and its role in the grammaticalization of definiteness in Slavic]. Bettina Bock and Maria Kozianka, eds. *Schleichers Erben: 200 Jahre Forschung zum Baltischen und Slavischen*. Hamburg: Baar-Verlag, 31–49.

Miličević Petrović, Maja, Teodora Vuković, Mirjana Mirić, Daria Konior, and Anastasia Escher. (2023) "Toward sociolinguistic corpora of Torlak". *Zeitschrift für Slavische Philologie* 79(1): 123–51.

Mladenova, Olga. (2007) *Definiteness in Bulgarian: Modelling the processes of language change*. Berlin/Boston: De Gruyter Mouton.

R Core Team. (2022) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rudin, Catherine. (2018) "Multiple determination in Bulgarian and Macedonian: An exploration of structure, usage, and meaning". Stephen M. Dickey and Mark Richard Lauersdorf, eds. *V zeleni drželi zeleni breg: Studies in honor of Marc L. Greenberg*. Bloomington, IN: Slavica Publishers, 263–86.

Stanković, Branimir. (2017) "DP and mandatory determiners in article-less Serbo-Croatian". *Acta linguistica academica* 64(2): 257–79. DOI 10.1556/2062.2017.64.2.5.

Stanojević, Marinko. (1911) "Severno-timočki dijalekat: Prilog dijalektologiji istočne Srbije" [The northern Timok dialect: A contribution to the dialectology of eastern Serbia]. *Srpski dijalektološki zbornik* 2: 360–463.

Stojanov, Stojan. (1983) *Gramatika na săvremennija bălgarski knižoven ezik* [Grammar of the contemporary Bulgarian literary language]. Vol. 2. Sofia: Bălgarskata akademija na naukite.

Šimko, Ivan. (2020) "Definiteness markers in the *Life of St. Petka*". *Zeitschrift für Slawistik* 65(2): 272–307.

Tomić, Olga Mišeska. (2006) *Balkan Sprachbund morpho-syntactic features*. Dordrecht: Springer. [Studies in Natural Language and Linguistic Theory, 67.] DOI 10.1007/1-4020-4488-7.

Topolinjska, Zuzanna. (2006) "Are there three variants of the definite article in Macedonian?". *Južnoslovenski filolog* 62: 7–15.

———. (2009) "The linear order of adjectival modifiers (AM) in the Macedonian and Bulgarian noun phrase (NP) (based on the analysis of standard Macedonian texts)". Mila Dimitrova-Vulchanova and Olga Mišeska

Tomić, eds. *Investigations in the Bulgarian and Macedonian nominal expression*. Trondheim: Tapir Academic Press, 51–73.

Vuković, Teodora. (2019) Torlak ReLDI Tagger 2019. Available at: https://github.com/bravethea/Torlak-ReLDI-Tagger-2019. Last accessed 3 August 2022.

———. (2021) "Representing variation in a spoken corpus of an endangered dialect: The case of Torlak". *Language resources & evaluation* 55: 731–56. DOI 10.1007/s10579-020-09522-4.

Vuković, Teodora, Anastasia Escher, and Barbara Sonnenhauser. (2022) "Degrees of non-standardness: Feature-based analysis of variation in a Torlak dialect corpus". *International journal of corpus linguistics* 27(2): 220–47. DOI 0.1075/ijcl.20014.vuk.

Vuković, Teodora, Mirjana Mirić, Anastasia Escher, Svetlana Ćirković, Maja Miličević Petrović, Andrey Sobolev, and Barbara Sonnenhauser. (2023) "Under the magnifying glass: Dimensions of variation in the contemporary Timok variety". *Zeitschrift für Slavische Philologie* 79(1): 153–94.

Vuković, Teodora and Tanja Samardžić. (2018) "Prostorna raspodela frekvencije postpozitivnog člana u timočkom govoru" [Spatial distribution of the frequency of the postpositive article in the Timok vernacular]. Svetlana Ćirković, ed. *Timok: Folkloristička i lingvistička terenska istraživanja 2015–2017*. Knjaževac: Narodna biblioteka "Njegoš", 181–200.

Zwicky, Arnold. (1977) *On clitics*. Bloomington, IN: Indiana University Linguistics Club.

Zwicky, Arnold and Geoffrey Pullum. (1983) "Cliticization versus inflection: English *n't*". *Language* 59(3): 502–13.

Teodora Vuković
Digital Society Initiative
and
Department of Computational Linguistics
University of Zurich
Zurich, Switzerland
teodora.vukovic2@uzh.ch