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): 245–69, 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.F.SG.ACC 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.ACC-DEM.ACC
'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.

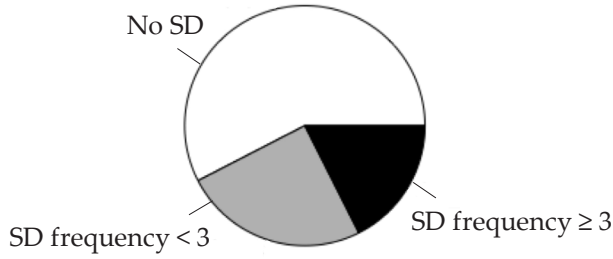


Figure 1. The distribution of SD frequency across speakers (per 1,000 tokens)

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 above.[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 ref-

erents 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 an

[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).

anaphoric interpretation. In example (6), the referent marked with an SD denotes a referent previously mentioned in the discourse, while also referring 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.ACC
'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

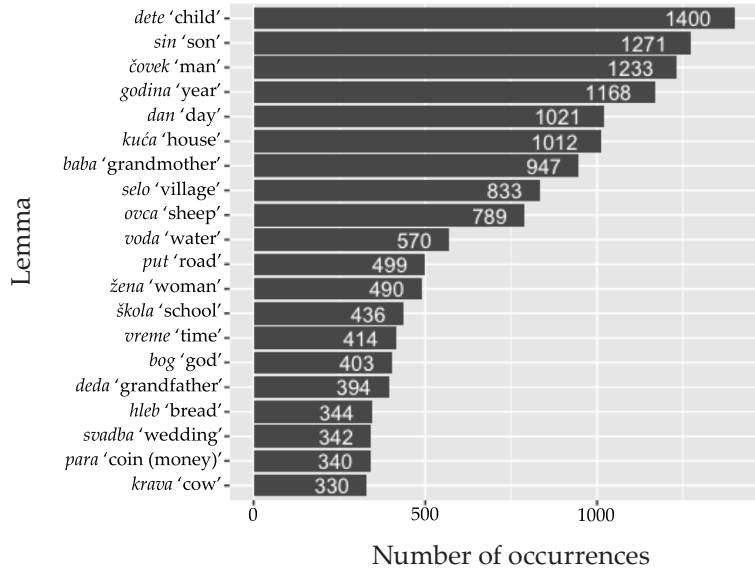
	Max (abs freq)	Mean (abs freq)	SD (abs freq)	CI LL	CI UL
Bare noun	1,400	5.48	33.35	97.50%	97.90%
Noun + SD	27	0.07	0.77	2.07%	2.52%

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

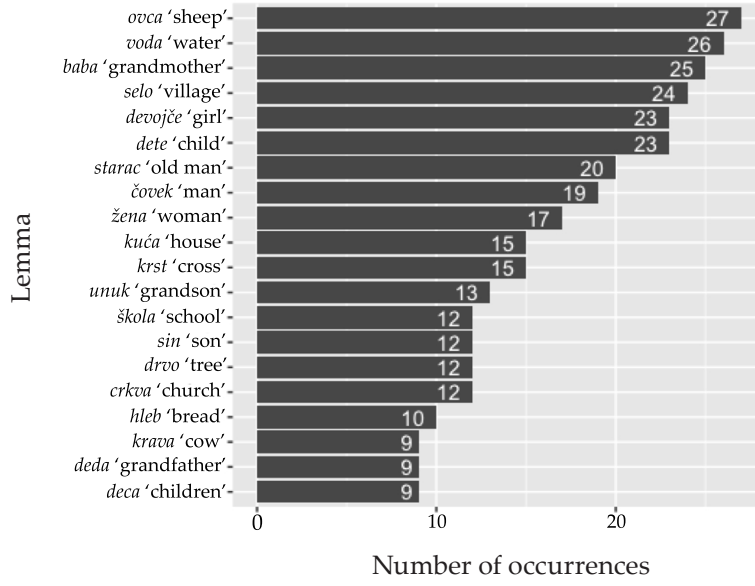
[14] Function "glm()" was used from the R package Stats (R Core Team 2022).

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



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 described in the results below.

When it comes to the analysis of the semantic criteria, both distinctions (count vs. mass and concrete vs. abstract) 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).

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

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, distribu-

tional patterns within the structure of the noun phrase can be used to assess their grammatical status. Their linear position and co-occurrence with other 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 con-

stituents 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.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 sam-

ple 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 morpho-

syntactic 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