



volume 31, numbers 1–2

2023

Journal of the Slavic Linguistics Society

Editors-in-Chief

Franc Marušič, *University of Nova Gorica*

Rok Žaucer, *University of Nova Gorica*

Associate Editors

Boban Arsenijević, *University of Graz*

Neil Bermel, *University of Sheffield*

Christina Y. Bethin, *Stony Brook University*

Wayles Browne, *Cornell University*

Barbara Citko, *University of Washington, Seattle*

Uwe Junghanns, *University of Göttingen*

Moreno Mitrović, *Humboldt University*

Roumyana Pancheva, *University of Southern California*

Catherine Rudin, *Wayne State College*

Eun-Ji Song, *Seoul National University*

Editorial Board

Stephen Dickey, *U. of Kansas*

Masako Ueda Fidler, *Brown U.*

Hana Filip, *Heinrich-Heine-U.*

George Fowler, *Indiana U.*

Egbert Fortuin, *Leiden U.*

Steven Franks, *Indiana U.*

Frank Y. Gladney, *U. of Illinois*

Laura Janda, *U. of Tromsø*

Keith Langston, *U. of Georgia*

James Lavine, *Bucknell U.*

Maria Polinsky, *U. of Maryland*

Ljiljana Progovac, *Wayne State U.*

Irina Sekerina, *College of Staten Is.*

Luka Szucsich, *Humboldt U.*

Markéta Ziková, *Masaryk U.*

Anton Zimmerling, *Pushkin Institute*

Jacek Witkoś, *Adam Mickiewicz U.*

Editors-in-Chief (and general editorial address)

Franc Marušič
Rok Žaucer
Center for Cognitive Science of Language
University of Nova Gorica
Vipavska 13, Rožna Dolina
SI 5000 Nova Gorica, Slovenia

[FM] ++386 5 3315 207
[RZ] ++386 5 3315 277
franc.marusic@ung.si
rok.zaucer@ung.si

Managing Editors

Jordan Hussey-Andersen
Ellen Tamura

journalofslaviclinguistics@gmail.com

Book Review Editor

Wayles Browne
Department of Linguistics
Cornell University
Ithaca, NY, USA

ewb2@cornell.edu

Associate Editors

Boban Arsenijević
Department of Slavic Studies
University of Graz
Graz, Austria

boban.arsenijevic@uni-graz.at

Neil H. Bermel
School of Languages and Cultures
University of Sheffield
Sheffield, United Kingdom

n.bermel@sheffield.ac.uk

Christina Y. Bethin
Department of Linguistics
Stony Brook University
Stony Brook, NY, USA

christina.bethin@stonybrook.edu

Barbara Citko
Department of Linguistics
University of Washington
Seattle, WA, USA

bcitko@u.washington.edu

Uwe Junghanns
Slavic Department
University of Göttingen
Göttingen, Germany

uwe.junghanns@phil.uni-goettingen.de

Moreno Mitrović
Department of Slavic and Hungarian Studies
Humboldt University
Berlin, Germany

moreno@bled.institute

Roumyana Pancheva
Department of Linguistics
University of Southern California
Los Angeles, CA, USA

pancheva@usc.edu

Catherine Rudin
Languages and Literatures Department
Wayne State College
Wayne, NE, USA

carudin@gmail.com

Eun-Ji Song
Department of Russian Language and Literature
College of Humanities, Seoul National University
Seoul, South Korea

eunjis@snu.ac.kr

Slavic Linguistics Society

The Slavic Linguistics Society is an organization devoted to the systematic and scholarly study of the Slavic languages. Membership in the Society is open to all linguists regardless of field, theory, framework, or school. Members are entitled to present papers at annual meetings and receive the *Journal of Slavic Linguistics*, which is the official journal of the Slavic Linguistics Society. Individuals receive a subscription to *JSL* with their annual membership in the Slavic Linguistics Society. To join, go to <http://www.slaviclinguistics.org>.

SLS is governed by an Executive Board consisting of six members elected at the annual meeting for three-year staggered terms. The Executive Board negotiates arrangements for the annual meetings with host institutions. Meetings typically alternate between North America and Europe.

Executive Board

Andrei Antonenko, *Stony Brook University*

Marco Biasio, *University of Pisa*

Wayles Browne, *Cornell University*

Jadranka Gvozdanović, *Heidelberg University*

Aida Talić, *University of Illinois Urbana-Champaign*

Jelena Živojinović, *University of Graz*

Guidelines for Contributors

JSL is intended to address issues in the description and analysis of Slavic languages of general interest to linguists, regardless of theoretical orientation. Papers dealing with any aspect of synchronic or diachronic Slavic phonetics, phonology, morphology, syntax, semantics, or pragmatics will be considered, as long as they raise substantive problems of broad theoretical concern or propose significant descriptive generalizations. Comparative studies and formal analyses are especially encouraged. Submissions ranging from full-length articles to shorter remarks and replies will be considered, along with substantive reviews.

Manuscripts should be submitted through the journal's website, where submission guidelines are available, as well as online versions of volumes starting with volume 19 (2011).

JSL's World Wide Web page: <http://ojs.ung.si/index.php/JSL>

The *Journal of Slavic Linguistics* publishes two regular issues per year, totaling approximately 360 pages: Winter–Spring and Summer–Fall. Each year *JSL* also publishes an extra open-access, online-only issue that contains the proceedings from the Formal Approaches to Slavic Linguistics (FASL) conference. For regular issues, library and institutional subscriptions are available from Slavica Publishers at the following rate:

Institutions \$60.00

Individuals should join the Slavic Linguistic Society (<http://www.slaviclinguistics.org>), which carries with it an annual subscription to *JSL*.

Airmail supplement for institutions outside of North America: \$11.00/year.

A limited number of back issues are available for \$24.00 apiece.

Indexing and Abstracting: *American Bibliography of Slavic and East European Studies*, ERIH (*European Reference Index for the Humanities and Social Sciences*), *Humanities International Index*, IBZ (*Internationale Bibliographie der Geistes- und Sozialwissenschaftlichen Zeitschriftenliteratur*), *MLA International Bibliography* (*Modern Language Association*), *OCLC ArticleFirst*, and *Web of Science Emerging Sources Citation Index*

Each contribution copyright © by its author(s). All rights reserved.
This issue was published in July 2024.

ISSN: 1068-2090

Technical Editors: Jordan Hussey-Andersen and Ellen Tamura

Cover design: Nikki Life

Slavica Publishers
Indiana University
1430 N. Willis Drive
Bloomington, Indiana USA
47404-2146



[Tel.] 1-812-856-4186
[Fax] 1-812-856-4187
[Toll-free] 1-877-SLAVICA
slavica@indiana.edu
<http://www.slavica.com>

From the Editors

vii

articles

Chloe Castle

Czech, Mate: Grammatical Replication and Shift
in South Australian Czech

1

Masako U. Fidler and Václav Cvrček

Zone-Flooding as a Discursive Strategy
of Czech Anti-System News Portals

61

Agnieszka Kaleta

The Semantics of Clausal Complementation:
Evidence from Polish

99

Ivana LaTerza, Petya Osenova, and Boban Karapejovski

Binding in South Slavic and DP: A Data-Driven Approach

133

Luca Molinari

The Syntax of Bulgarian *edin* ‘one’

161

Tore Nessel and Anastasia Makarova

Threatening in Russian with or without *sja*: *Grozit’* vs. *grozit’sja*

215

Teodora Vuković

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

245

reviews

Ronelle Alexander

Senahid Halilović (chief of the project), Mehmed Kardaš,
Amela Ljevo-Ovčina, and Emira Mešanović-Meša.

Bosanskohercegovački lingvistički atlas I: Fonetika

[Bosnian-Herzegovinian linguistic atlas I: Phonetics]

271

John Dunn	
Thomas Rosén. <i>Russian in the 1740s</i>	275
Victor Friedman	
Anna-Maria Sonnemann. <i>Language Families in Contact: The Mutual Impact of Slavic and Romani</i>	279
Grant Lundberg	
Ágoston Pável. (Marc L. Greenberg, ed./trans.) <i>Prekmurje Slovene Grammar: Avgust Pavel's Vend Nyelvtan (1942)</i>	285
Anita Peti-Stantić	
Tomasz Kamusella. <i>Politics and the Slavic Languages</i>	287
Donal F. Reindl	
Katarzyna Bednarska, Dorota Kruk, Borislav Popov, Olga Saprikina, Traci Speed, Kamil Szafraniec, Svitlana Terekhova, Radislav Tsonev, and Aneta Wysocka, eds. <i>Contributions to the 23rd Annual Scientific Conference of the Association of Slavists (Polyslav)</i>	293
Robert Rothstein	
Jan Fellerer. <i>Urban Multilingualism in East-Central Europe: The Polish Dialect of Late Habsburg Lviv</i>	301
Danko Šipka	
Predrag Piper, Ivan Klajn, and Rajna Dragićević. <i>Normativna gramatika srpskoga jezika</i> [Normative grammar of the Serbian language]	303
Danko Šipka	
T. I. Vendina. <i>Praslovjanskoe slovo vo vremeni i prostranstve Slavii</i> [Proto-Slavic words in time and space of the Slavia]	307
Aida Talić	
Steven Franks. <i>Microvariation in the South Slavic Noun Phrase</i>	311
Anton Zimmerling	
Zrinka Kolaković, Edyta Jurkiewicz-Rohrbacher, Björn Hansen, Dušica Filipović Đurđević, and Nataša Fritz. <i>Clitics in the wild: Empirical studies on the microvariation of the pronominal, reflexive and verbal clitics in Bosnian, Croatian, and Serbian</i>	319

From the Editors

This is the entire volume 31, a double issue combining what would normally have been 31.1 and 31.2. We decided to combine 31.1 and 31.2 into a double issue in an attempt to reduce the overall delay of the volume (which is JSL's 2023 volume).

In the foreword to 30.2, we announced having taken the first concrete steps in JSL's attempt to become open access: starting to make new regular issues available on JSL's website in delayed open access (one year after publication) and making extra issues available on JSL's website in immediate open access. We now add that we have also started to make back issues available on JSL's website, with the website currently containing back issues all the way to volume 19 (2011).

The issue also marks a change in the editorial team: we welcome Ellen Tamura, who is replacing Jordan Hussey-Andersen as our Managing and Technical Editor.

Franc Marušič and Rok Žaucer
University of Nova Gorica
franc.marusic@ung.si & rok.zaucer@ung.si

Czech, Mate: Grammatical Replication and Shift in South Australian Czech

Chloe Castle

Abstract: Historical linguistics aims to investigate the innovation stage of a grammatical variant as well as the later community-wide propagation in order to fully understand the change (Fischer 2004). This paper focuses on individual contact-based grammatical innovations in a community setting, viewing the speaker as the “locus of change” (Weinreich 1953/1968: 1; Romaine 2005; Wei 2013). This provides a window into the types of innovations community members produce in a situation of shift, wherein such innovations may never become complete changes. The community studied in this article is the Czech South Australian community, whose language situation is previously unstudied. Utilizing Thomason’s (2001) steps for proving whether contact-induced structural change has occurred, this paper identifies several instances of possible grammatical “replication” innovations in the speech of individuals in this community (Heine and Kuteva 2005, 2008: 2; Kuteva 2017), as well as the influence of shift driven by “divergent attainment” (Polinsky 2018: 18) and intergenerational attrition. This is supported by findings of significant authors in the tradition of Czech diasporic linguistic research (Henzl 1982; Vašek 1996; Dutková 1998; Dutková-Cope 2001a, 2001b; Zajícová 2009, 2012). It is suggested here that the features found are possibly the result of shift and attrition processes and contact-induced language transfer acting together within a Dynamic System (Herdina and Jessner 2002).

1. Introduction

In this study, I investigate grammatical features occurring in the speech of ten individuals from the Czech South Australian community, particularly those representing **grammatical replication** and **borrowing** (Heine and Kuteva 2005; Kuteva 2017). Grammatical replication is a kind of transfer that does not involve phonetic substance of any kind, including contact-induced grammaticalization, restructuring, rearrangement, and loss (Heine and Kuteva 2003, 2005, 2008, 2010). Borrowing, on the other hand, is “reserved for transfers involving phonetic material, either on its own or combined with meaning” (Heine and Kuteva 2010: 86). This community is undergoing attrition and language shift, which are also key considerations in the analysis. I utilize a methodology of qualitative analysis of grammatical features drawn from authentic

speech, in alignment with the tradition of Czech diasporic linguistic research (Henzl 1982; Vašek 1996; Dutková 1998; Dutkova-Cope 2001a, 2001b; Zajíčová 2009, 2012). I support this with the use of Thomason's (2001: 93–94) steps for identifying contact-induced structural change and the dynamic model of multilingualism (Herdina and Jessner 2002). Thomason's steps particularly aid in detecting whether the features are instances of grammatical replication or are attrition- and shift-based. The dynamic model of multilingualism then offers possible insight into how these sources interact.

In this paper, I adopt the epistemological stance of Matras and Sakel (2007) in positing that a community-wide change begins at the level of an innovation by an individual speaker. Indeed, Fischer (2004: 10) suggests that the innovation stage of a grammatical change must also be investigated to provide a full understanding of "the system of grammar with which adults innovate". Thomason (2014: 202) states that "any innovation ... is a potential language change—even a one-time speech error or a joking coinage—[and] the fate of every innovation is determined by a combination of linguistic and (especially) social factors". In line with these scholars, this research centers the individual as the "locus of change" (Weinreich 1953/1968: 1; Romaine 2005; Wei 2013) and analyzes innovations created by individuals in this community setting. This loosely follows Clyne's (2003: 96) approach in considering "change" in contact situations for individuals rather than for an entire speech community. Using this approach allows for an understanding of the bilingual grammatical features occurring **in the individual**, and how community members utilize the grammatical resources available to them. The innovative grammatical features found are thus labeled as unconventionalities (Doğruöz and Backus 2009): unconventional speech productions that may not necessarily result in propagation and community-wide change.

The Czech diaspora is a minority among minorities in the Australian linguistic landscape, and therefore not a key focus in prominent works on the linguistic tapestry of Australia (Clyne 2003; Clyne and Kipp 1996, 2006). Languages that are, or were, more widely spoken in Australia are at the center of such analyses: German, Dutch, Croatian (Hlavac 2000), Vietnamese (Ho-Dac 1996, 2003), and more. It is important to deepen understanding of the many language communities in Australia in order to better support them in language maintenance (if this is their desire), and to express and support the validity of these community members' languages. I aim to record and contribute to the information available on Australian community languages: the ways that they are used, considered, and how language contact and attrition processes have played a role in linguistic outcomes.

This paper considers both intergenerational language attrition (also called shift), wherein subsequent generations have reduced input and therefore divergent attainment, and intragenerational language attrition. It is recognized that there is an influence of language contact **within** the attrition process

(Preston 1982; Andersen 1982; Sharwood Smith 1989; Seliger and Vago 1991; Huffines 1991; Sharwood Smith and van Buren 1991; Polinsky 1997; Altenberg 2010). The paper compares the linguistic outcomes of this community with other Czech diasporic communities studied in the past and considers how the level of technology available to speakers in different times and other factors may influence language attrition.

Section 2 provides a background to the study, introducing the South Australian Czech community and the relevant findings from other Czech diasporic communities. Section 3 explains the method, including the data gathering and coding processes. Section 4.1 shows the results of the study and interacts with the literature in providing a qualitative analysis of the observed features. Section 4.2 provides an analysis in terms of Thomason's (2001: 93–94) steps for establishing structural interference in a receiving language and the dynamic theory of multilingualism. Section 5 concludes the paper, presenting an overall summary, limitations, and future research possibilities.

2. Background

In this section, the background of the speech community is explored and grammatical borrowing in other Czech diasporic situations is considered.

2.1. Who Are the Czech South Australians?

Czech immigration into the state of South Australia coincides with key events within Czech history. There was some Czech immigration to Australia before WWII, but the major waves of immigration occurred in 1949 following the 1948 communist takeover of Czechoslovakia, in the 1970s following the 1968 Prague Spring¹, and after the Velvet Revolution in 1989 (Vaculík 2009; Brouček et al. 2019).

In the first wave, 1,500 Czechs arrived in South Australia (SA), many of whom had previously migrated to Germany after fleeing Czechoslovakia (Migration Museum 2020). These people were generally not welcomed by those who had come pre-WWII due to political views, e.g., the Sydney expatriate circle was operated by communists at the time (Vaculík 2009). This new group of immigrants thus formed “reactionary” sporting and social clubs as community refuges (Vaculík 2009: 242–44). The Czechoslovak Club was formed in

¹ The Prague Spring was a period of liberalization in Czechoslovakia wherein many reforms occurred, including greater freedom of expression for the press and loosening of restrictions on travel, granted by Alexander Dubček, who became first secretary of the Czechoslovak Communist Party on 5 January 1968. This period ended on 21 August 1968, when the Warsaw Pact forces invaded and occupied the country, and the reforms were purged the following year.

1949 and was incorporated as an official body in the 1950s (Migration Museum 2020). In the second major wave, around 1,000 Czechs settled in SA, and these political refugees were aided by the Central Committee of the Czechoslovak Compatriots Association in Australia and New Zealand² in conjunction with the Australian government (Vaculík 2009). From the mid-1990s onwards, following the 1989 Velvet Revolution, many Czechs have migrated to Australia and New Zealand for personal and professional reasons (Brouček et al. 2019).

The differences in time of arrival affect the national and sociolinguistic identities and attitudes of the Czech Australians in many ways. The timeline of Czech immigration interacts with the government policies and community attitudes in Australia at each time (Clyne and Kipp 2006). The official policy of the Australian government remained assimilationist and hostile toward the maintenance of distinct sociocultural identities of immigrant groups until the 1970s (e.g., the White Australia policy³; Clyne and Kipp 2006). In the post-1970 period, multiculturalism and government support arose to create a context more conducive to promoting and maintaining sociocultural identity in immigrant groups (Clyne and Kipp 2006). The role of government policies and dominant community attitudes towards the presence of ethnic languages are an important factor in language maintenance or language shift (Pauwels 1988; Clyne and Kipp 1996).

The Czechoslovak Club in SA, Inc., or *Československý klub v Jižní Austrálii*, is the sole Czech and Slovak club existing in South Australia today,⁴ and it served as the fieldwork location for this research. The Club was established in 1949 and incorporated as an official body in the early 1950s (Migration Museum 2020). Its premises is a hall located in the suburb of Brompton, 6.2km north of the city of Adelaide's Central Business District (CBD), South Australia. It is attended by more recent arrivals and older generations (and their children and grandchildren) alike. The Club has an aim to "connect all Czechs and Slovaks from South Australia in a strong community that keeps and promotes national ideas based on united friendship and mutually honest social

² *Ústřední výbor krajanské Československý Asociace v Austrálii a na Novém Zélandu*: this association no longer exists, but there are currently 15 Czech and Slovak community associations in Australia and New Zealand (Embassy of the Czech Republic in Canberra 2021).

³ This was a series of policies restricting immigration of non-white, non-British groups to Australia from 1901. The policies were fully dismantled in 1973. During this time, racist anti-immigration propaganda was rampant and assimilation to the majority British population was strongly encouraged (National Museum Australia 2021). Post-WWII, the government allowed more non-British white immigrants into Australia, but government policy (and public opinion) stated that migrants should assimilate (Migration Heritage Centre 2010).

⁴ There is, however, a separate Slovak Club.

relations” (Charles Sturt Council 2019). The Club is the center of Czech cultural life in the region, providing weekly dinners, social and cultural events such as St. Mikuláš Day and the anniversary of the declaration of Czechoslovak independence, welfare services, and once-weekly children’s language classes. Whilst Czech is used at the Club, it is not used by all and tends to depend on the individual’s generation (Castle 2021). There are approximately 280 Club members, though of these, I observed approximately 50–60 key active members at the events attended, including the Annual General Meeting, the Christmas wreath-making event, and several Club dinners. At the time of the 2016 census, there were 473 Czech-born South Australians and 1,679 South Australians of Czech descent,⁵ 0.02% and 0.1% of the South Australian population, respectively (Australian Bureau of Statistics 2017, 2022). There are many more South Australians of Czech descent and Czech-born South Australians than there are Club members, suggesting a somewhat scattered, wider Czech South Australian community with a tighter-knit Club community at its center. The existence of a scattered, wider community is corroborated by the spread of those who reported Czech ancestry on the 2016 census throughout the Greater Adelaide region (Figure 1).⁶ There is a slight concentration of those with Czech ancestry in the northeastern suburbs. As Brompton is to the north of the city, the Club may be more frequented by those living in the north. The Adelaide CBD and Brompton are demarcated on each figure.

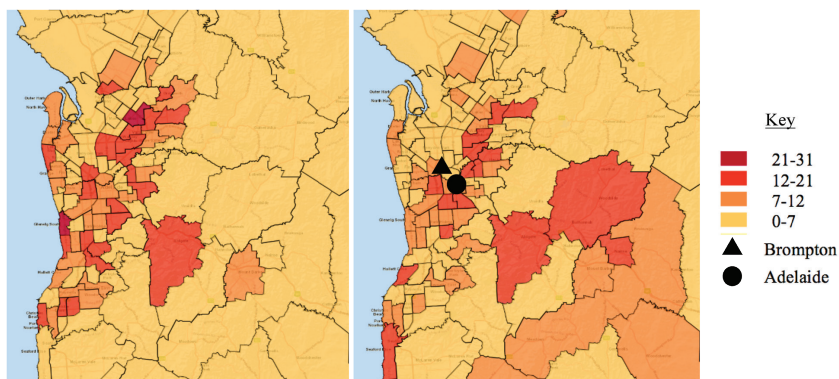


Figure 1. People who reported Czech ancestry as their first option (left) and second option (right) (Australian Bureau of Statistics 2022)

⁵ Census statistics were retrieved using the TableBuilder tool (Australian Bureau of Statistics 2022). The original tables and instructions for generating them are provided in Appendix 1 and 2.

⁶ The two maps presented in Figure 1 were created using the TableBuilder tool (Australian Bureau of Statistics 2022). The original statistics and instructions for generating them are provided in Appendix 3 and 4.

High-shift groups, in terms of language loss, tend to be those with a relatively smaller cultural distance from the dominant group, i.e., in terms of religion, historical consciousness, culture, and a lack of taboo around exogamy (Clyne and Kipp 1996, 2006). Hailing from a Central European nation, Czechs are culturally different from Anglo-Australians (the dominant group in this case), but not dramatically so, and exogamy is not frowned upon. Other factors affecting shift or language maintenance are whether language is a core value for the individual and community, and the length of residence and socio-political factors in the homeland and in Australia (Stoessel 2002; Clyne and Kipp 2006). Victoria and South Australia have had relatively lower shift rates than other Australian states for European languages, which can be partially attributed to a tradition of multicultural policies in these places⁷ (Clyne 1982; Clyne and Kipp 1996).

2.2. Grammatical Changes in Other Diasporic Czech Communities

There have been several studies of language change in minority Czech communities elsewhere in the world: in Texas (Dutková 1998; Dutkova-Cope 2001a, 2001b; Eckert 2006; Pintová 2009; Eckert and Hannan 2009; Eckertová 2017); Chicago (Rakusan 1993); America in general (Henzl 1982; Vašek 1996); and Paraguay (Zajícová 2009). Czech immigration into Texas and the wider US occurred in the mid-19th century, and into Paraguay from 1927–1939 (Pintová 2009; Vašek 1996: 71; Zajícová 2012). The general picture that emerges from this research is that when languages first come into contact, and for the generation following, grammatical changes do not necessarily involve wholesale simplification, but rather tend to involve structural convergence between Czech structures and those of the majority language (Dutkova-Cope 2001b; Zajícová 2012). However, as the younger generations experience divergent attainment, their Czech begins to simplify and structural relations are lost (Dutková 1998; Zajícová 2012).

The grammatical features found in those communities which are relevant to the findings in this paper are presented in Table 1. The communities wherein the same features occur are ticked. The two language groups display almost all of the same features. This, coupled with the fact that the contact languages are not only English and Czech but also Spanish and Czech, leads to the suggestion that communities do not have their own individual paths of development in terms of language shift and maintenance. Rather, there are commonalities in the developmental path of Czech, regardless of the contact language. The changes may thus be more typical of Czech in a contact situation, i.e., possibly accelerating already existing slow changes in the language,

⁷ Australia is a federation and therefore states are vested with legislative power over areas including education, community services, and health.

or undergoing particular types of changes under attrition conditions. This comparison and information informs the study as to which features are typical of a contact situation involving Czech.

Table 1. Grammatical features in Czech diaspora communities

Grammatical feature	American Czech ⁸	Paraguayan Czech
Overt subject marking	✓	✓
Preposition instability	✓	✓
Loss of case distinction	✓	✓
Loss of gender distinction	✓	✓
Reflexive pronoun instability	✓	✓
Increasingly analytic syntax	✓	
Tentative article formation	✓	✓

Now that the community socio-historical background and the grammatical features occurring in other similar diasporic communities have been established, I move on to discuss the methods adopted for this study.

3. Method

In this section, I discuss the data collection, participant information, and data coding and analysis.

3.1. Data Collection

The study involves four observation sessions with groups of two to three people in the Adelaide Czechoslovak Club, and six semi-structured interviews conducted in English.⁹ The participants in the observation sessions were both video- and audio-recorded. To prompt conversation, participants were given

⁸ “American Czech” does not refer to Czech spoken by a single diaspora community but is an amalgamation of the phenomena found in various communities in different time periods across the US (thus involving language contact with English), including Texas (Dutková 1998; Dutkova-Cope 2001a, 2001b; Eckert 2006; Pintová 2009; Eckert and Hannan 2009; Eckertová 2017), Chicago (Rakusan 1993), and America in general (Henzl 1982; Vašek 1996).

⁹ Only six of the ten participants were available for the subsequent interviews.

discussion sheets written in Czech only (Appendix 5). The discussion sheet included topics such as family, life memories, and the upcoming Christmas festivities to encourage speakers to speak more naturally, as speakers are more likely to approximate their casual style when they become emotionally involved in the narration (Labov 1972). The participants were thus engaged in relaxed, everyday discourse.

I did not participate in the discussion, so as to avoid the possibility of participant accommodation to my lower level of fluency in Czech. However, I was present but seated away from the participants, in the corner of the room. After recording, I transcribed the participant discussions using ELAN. A native Czech-speaking transcriber from an external company¹⁰ completed a second transcription to ensure that it was correct.

The sample is non-random: it is shaped through referrals biased towards those perceived as having adequate bilingual abilities by community members. Sampling is skewed towards females as referrals from the female club manager tended to favor female speakers. However, this does not necessarily represent an issue and could in fact be helpful to the study, considering that women are generally the innovators in linguistic change (Labov 1990). As with Dutková's study, "practical considerations partly dictate[d] sample size" (Dutková 1998: 93; a similar point is made in Milroy 1987: 23). However, I endeavored to obtain a sample with a varied age range, speaker ability, ancestral regions, and educational levels to maximize the chance of finding different features amongst a relatively small participant group, as displayed in Table 2. The duration of the sessions is given in Appendix 7.

The questionnaire includes the number of years residing in Australia to avoid situations where newly arrived Czechs with a possible lower competency in English would skew the dataset.

Individuals were required to have adequate proficiency in both languages, which is determined with a self-test (Table 3 on p. 10),¹¹ as well as a content analysis of the observation sessions and sociolinguistic interviews for information on social networks (Table 4 on p. 11).¹²

¹⁰ The company is called Knockhundred Translations.

¹¹ This self-test questionnaire can be found in Appendix 6.

¹² Content analysis "contextualises questionnaire reports ... more generally allowing for [their] interpretation" (Torres Cacoullos and Travis 2018: 58). Content analysis involves the processing and coding of qualitative information (e.g., in this case, information about frequency of language use, people with whom participants use the language, etc.).

Table 2. Participant variables

Name¹³	Age	Gender	Years in Australia	Generation	Educational level	Region of origin
Adéla	>50	F	>20	1.5 ¹⁴	Vocational education	Bohemia
Dana	<50	F	>20	1.5	Bachelors	Moravia
Eva	<50	F	>20	2nd	Bachelors	Australian born
Jana	>50	F	>20	2nd	Bachelors	Australian born
Ivana	<50	F	10–20	1st	Bachelors	Bohemia
Kamila	>50	F	10–20	1st	Masters	Moravia
Milada	<50	F	>20	1st	Bachelors	Bohemia
Zuzana	<50	F	10–20	1st	Masters	Moravia
Roman	>50	M	>20	1st	High school	Bohemia
Martin	<50	M	10–20	1st	Masters	Bohemia

In all cases except for Ivana and Roman, participants' better language reflects their generation: all 1.5- and 2nd-generation Czech South Australians have English as their better language, while all 1st-generation Czech South Australians have Czech as their better language.

It is important to consider the social networks (Milroy 1987) of the participants as this reflects the language(s) that are most commonly used by them and therefore the languages that are most well maintained (Stoessel 2002).

In the participant information sheet,¹⁵ I informed participants that the study was about communication in the Czech community in South Australia. I stated that the project involves analyzing how bilingual Czech Australians

¹³ Names have been changed for purposes of confidentiality.

¹⁴ In this study, the generations are split into three groups: 1st generation, 1.5 generation, and 2nd generation. The 1.5 generation refers to those individuals who were born in the heritage country (here, Czech Republic) but moved to the new country (here, Australia) in childhood with their parents (Rumbaut and Ima 1988; Rumbaut 1994, 1997, 2004).

¹⁵ This form was provided to potential participants to gain an understanding of what the study is about, what they are invited to do, the length and benefits of the project, and how their information will be used.

Table 3. Participant language proficiency and generation

Generation	1st					1.5			2nd	
	Ivana	Kamila	Milada	Zuzana	Roman	Martin	Adéla	Dana	Eva	Jana
Self-score (E)	9 =	7	7	9	10 =	9	10 ✓	10 ✓	10 ✓	10 ✓
Self-score (C)	9 =	10 ✓	8 ✓	10 ✓	10 =	10 ✓	7	5	8	3

Key: ✓ more proficient in this language
= equal
E English
C Czech

Table 4. Participant social networks and use of languages

Participant	Network information
Adéla	Speaks Czech with Czech husband and at the Czech Club. Used to speak English at work but has now retired. Currently uses Czech more often, with Czech friends (both over the phone and in Adelaide).
Dana	Mixes Czech and English at the Club. Occasionally speaks Czech with her children but mostly speaks English. Speaks English with partner. Speaks Czech or mixes Czech and English with her mother.
Eva	Speaks English with her Australian husband and children. Mixes Czech and English at the Club. Speaks English when in front of others (or Czech more quietly). Uses Czech only with Czech relatives (on the phone or when travelling). Mixes Czech and English with her parents.
Jana	Speaks English with her Australian husband. Mixes Czech and English or Czech only at the Club. Travels to the Czech Republic every few years.
Ivana	Speaks English with children when in front of others, but Czech when alone with them. Speaks English with Australian husband.
Kamila	Speaks Czech with her son in Australia (when daughter-in-law not present) and son in Czech Republic. Speaks Czech in the Club.
Milada	Speaks Czech with Czech husband. Speaks Czech at the Club. *
Zuzana	Speaks English with Australian husband. Speaks Czech at the Club. Travels to the Czech Republic every year.*
Roman	Speaks Czech at the Club.*
Martin	Speaks Czech at the Club.*

* Insufficient further information (did not participate in interview)

converse with one another in Australia as a result of language contact. I did not provide information beyond this (i.e., that I was focusing on grammar), in order to avoid excessive self-monitoring of grammar and therefore potentially fewer borrowing events. I encouraged them to use Czech but to speak as naturally as possible, even if that includes some English.¹⁶

3.2. Data Coding and Analysis

Instances of potential borrowing (morphological transfer) and grammatical replication were identified by myself and two Czech research assistants from Palacký University Olomouc. Previous Czech diasporic studies were used as an approximate guide as to what features may be found (whilst also analyzing for other features), and assistants were instructed to highlight phenomena that sounded unusual to them. Each assistant aimed to analyze different phenomena in their assessments to increase the richness of the results found. Assistant 1 focused on syntax, while Assistant 2 focused on morphology and subject-verb agreement.

It is recognized here that Czech is a unique, “intralinguistic”¹⁷ diglossic language situation (Bermel 2000: 34). There is a standard literary variety used in formal situations and in writing (*spisovná čeština*),¹⁸ and an unofficial variety used in speech (*obecná čeština*, or Common Czech) (Bermel 2000). It differs from other classic diglossic situations in that there is no portion of the community that uses the standard language as an L1, and there is not enough of a difference between the codes for the boundaries between them to be clearly marked (Bermel 2000). Bermel (2000: 34) states that, as Common Czech (CC) is not defined or codified in any official manner, “the only arbiters [of CC] are native speakers, preferably ... educated ones from certain parts of the Czech Republic”. The research assistants had access to both the video and audio recordings, as well as the transcripts and metadata, so that they could socially gauge the expected variety, both in terms of social context and participants’

¹⁶ This study was approved by the Adelaide University Ethics Committee (Approval No. H-2018-230).

¹⁷ The two varieties discussed here share enough syntax, morphology, phonology, and vocabulary that “many utterances cannot clearly be assigned to one or the other variety” (Bermel 2000: 16).

¹⁸ This is not a typical case of a written variety which has emerged from a spoken variety: it was purposefully developed during the National Revival of the 19th century (Bermel 2000). Leading intellectuals chose to draw on the “‘golden age’ of Czech prose: the era of the Kralice Bible” (the late 1500s) (Bermel 2000: 12). For more on this, see Bermel 2000 and Wilson 2008.

region of origin.¹⁹ They were therefore able to keep the context in mind when assessing whether the speech data sounded unnatural to them.

The Czech National Corpus (CNC)²⁰ is also utilized in the analysis, using the KonText application (Machálek 2014) for searching attestations and the Word at a Glance²¹ application (Machálek 2019) to indicate frequency of pronoun use over the years (see §4.1.1). The results of the attestation search supplement the qualitative analysis of each example and are available in Appendix 8. It is recognized that the corpus does not always allow for an understanding of the pragmatic context of the situation. However, there is a precedent in the Czech language contact literature, which this paper aims to follow, of using the data collected and making comparisons with other varieties of Czech to make calculated speculations on the phenomena occurring (Henzl 1982; Vašek 1996; Dutková 1998; Zajícová 2009, 2012).

4. Results and Discussion

In this section, the grammatical features found are discussed in detail, and a summary is given. Further analysis using Thomason's (2001) framework and the Dynamic Theory of Multilingualism (Herdina and Jessner 2002) is provided.

4.1. What Grammatical Features Were Found?

The Czech South Australian participants utilized the grammatical resources available to them in a variety of different ways.

4.1.1. Overt Subject Marking in pro-Drop Czech

Czech is a pro-drop language. However, the subject pronoun is included with the verb for the discourse-pragmatic purpose of emphasis (Zajícová 2009).

¹⁹ Several participants were from Moravia, a fact that was viewable to the research assistants in the metadata. As the research assistants were studying in Olomouc (located in Moravia), it is possible that they were alert for Moravian features as well.

²⁰ The Czech National Corpus is comprised of different subcorpora which together provide access to more than three billion words. There are different versions of certain corpora available. For searching the corpora, there are several applications available, including KonText, Slovo v kostce (Word at a Glance), and SyD. Where an application is first mentioned, its creators are cited. Where a (sub)corpus version is first mentioned, its creators are cited. All corpora assessing Czech in this study are from the CNC.

²¹ This interface allows for a comparison of frequency of use over the period 1998–2017.

The overt subject pronoun also occurs more frequently in colloquial²² speech (Janda and Townsend 2000).

An example of the emphasized subject is shown in (1) below (throughout the examples, the focus of interest is highlighted in bold):

(1) a. Standard Czech

Už jsme spolu mluvíli.
already AUX.1PL together spoke²³
'We've already talked together.'

b. Emphasized subject/colloquial

My jsme už spolu mluvíli.
we AUX.1PL already together spoke
'We've already talked together.'

Use of the overt pronoun varied across participants. There were numerous instances of overt pronoun usage which sounded unnatural to the research assistants in the social circumstances of each discussion (see Table 5).

It is evident from the data that while there are instances of subject pronoun use consistent with varieties of Czech used in the Czech Republic,²⁴ there are also instances in these participants' speech where use of subject pronouns would be unconventional.

In assessing which instances of the subject pronoun were relevant for this analysis, I implemented the following rules:

- If participants have used a pronoun coreferentially with a verb, it is included. However, if they have used a stand-alone pronoun, it is omitted in the analysis.
- The 3SG copula/dummy subject *to* 'it' is omitted in alignment with Torres Cacoullos and Travis's (2018: 139) choice to only include

²² Colloquial speech is here assumed to mean what Bermel (2000) calls "Common Czech". Janda and Townsend (2000: 4) directly contrast their "Colloquial Czech" with Literary Czech, so we can presume that this was what was meant here. In the Czech literature, there are some that argue for a separate category labeled "Colloquial Czech" (Kopečný 1949; Bělič 1959, 1960), which acts as an intermediate zone between Standard and Common Czech (Auty 1976). For more on this, see Wilson 2008.

²³ The abbreviations used in the glosses to denote grammatical information are in alignment with the Leipzig Glossing Rules.

¹⁸ This is consistent with Bermel's (2000: 20) suggestion that subject pronouns are used regularly in Common Czech (*obecná čeština*).

Table 5. Overt pronoun use per participant

Participant	Opportunities for subject pronouns to be used	No subject pronoun used	Subject pronoun used	Unconventional subject pronoun used
Adéla	70	31	39	11
Dana	83	45	38	12
Eva	61	40	21	3
Jana	102	87	15	3
Ivana	35	28	7	5
Kamila	88	60	28	13
Milada	38	29	9	2
Zuzana	133	79	54	8
Roman	31	26	5	5
Martin	34	28	6	1

human-specific subjects (though they chose to include only human-specific 3SG subjects, which I do not do here).

- Lexical pronouns are omitted (Torres Cacoullos and Travis 2018: 138).
- Instances of repetition are removed, including instances of switching between the polite and casual pronouns *ty* ‘you.SG’ and *vy* ‘you.PL’.

Two examples of participants’ selection of the unconventional subject pronoun in South Australian Czech are shown in (2) and (3) below:

(2) Zuzana

my jsme si to projeli, my se podíváme
 we AUX.1PL REFL it go.through.PST.PL we REFL look.PRF.1PL
 ‘we’ve gone through it, we’ll see’

(3) Zuzana

já musím jet domů
 I must.1SG go home
 ‘I have to go home’

Extensive use of the overt pronoun is not predictable from generational status or level of language proficiency; it is entirely possible that it is an individual stylistic choice. It is also possible that participants exhibit unconventional use of the subject pronoun due to the influence of English. Their use of the subject pronoun could be increasing the analytic nature of the language, a common outcome of language attrition (Andersen 1982; Maher 1991; Polinsky 1997). Andersen (1982: 83–100) outlines a general compensatory strategy employed by language users that involves using “free morphemes whenever possible, strung together linearly ... to express your meaning”, thereby leading to increased analyticity, regardless of whether the language (in this case, Czech) would normally use them. The general presence of English world-wide and the accompanying increasing exposure of Czech speakers to English in general may also be accelerating an increase in the use of the subject pronoun in Czech that can be seen in data from the Czech National Corpus²⁵ (SYN 7; Křen et al. 2018), as shown in Table 6.

²⁵ This is purely data on the use of subject pronouns in all contexts; the researcher does not have the resources available to make distinctions based on discourse-pragmatic or syntactic placement at this point. However, a generalized, non-context-dependent increase may still indicate that an increase is occurring in the pre-verbal context (the context analyzed in this paper).

Table 6. Pronoun use in the SYN 7 corpus in 1998 and 2017

Pronouns	1998 use ²⁶	2017 use
já	1,210.24 ~ 1,229.21	1,743.29 ~ 1,769.18
ty	55.21 ~ 59.33	68.44 ~ 73.65
on/ona/ono	3,372.17 ~ 3,403.74	4,567.24 ~ 4,609.04
my	116.4 ~ 122.34	159.04 ~ 166.94
vy	305.47 ~ 315.04	583.42 ~ 598.44
oni	2,021.29 ~ 2,045.76	2,502.53 ~ 2,533.53

While it is possible that the research assistants use the subject pronoun less in their varieties of Czech, it is important to note the plausibility of the claim that this feature occurs due to contact-induced transfer with English. It is also attested in Zajícová's (2009) study of Czech use in Paraguay, where she attributes likely causation to the joint influence of (internal) attrition processes and Spanish (contact-induced transfer).

4.1.2. Preposition Instability: Use, Non-Use, and Misuse

In Czech, certain prepositions are generally required in specific circumstances/syntactic constructions, which then require a particular case ending. Some examples of the case requirements for each preposition are as follows: *bez* 'without' (+ genitive case), *pro* 'for' (+ accusative case), and *s/se* 'with' (+ instrumental case). Table 7 on the following page shows the frequencies for this feature.

Adéla produced a grammatically unnecessary preposition in front of the adverb *tam* 'there', possibly modeled on the parallel English preposition, as shown in (4):

(4) Adéla

takže místo tu rodinu v tam
 so place.NOM DEM.ACC family.ACC in there
 'so instead of [in place of] the family **in** there'

²⁶ The corpus provides the lower and upper bounds of the estimated trend per million words, hence why ranges are presented in this table.

Table 7. Unconventional preposition use per participant

Participant	Total preposition possibilities	Conventional preposition used	Unconventional preposition used		
			Not included where conventional	Used where conventional	Unconventional choice of preposition
Adéla	36	34	0	1	1
Dana	37	35	1	0	1
Eva	25	25	0	0	0
Jana	48	45	0	0	3
Ivana	44	44	0	0	0
Kamila	36	36	0	0	0
Milada	30	30	0	0	0
Zuzana	55	55	0	0	0
Roman	24	24	0	0	0
Martin	24	24	0	0	0

One participant spoke without using a preposition, which is shown in (5):

(5) Dana

osobní třídy Ø čtvrtek-Ø
 personal classes ? Thursday-ACC?NOM?
 'personal classes on Thursday'

This example is interesting because whenever days of the week are discussed in this way, they require the preposition *v/ve*, in this case meaning 'on', which triggers the accusative case. However, the accusative case ending for *čtvrtek* is unmarked (i.e., the same as for the nominative case). It is difficult to tell whether the participant intended the noun to be in the accusative case. However, as this is only attested once, it may be the case that it represents a slip-of-the-tongue speech error as opposed to being an indicator of preposition drop. In any case, as English would also require a preposition in this example, it cannot be attributed to English influence.

Some participants utilized unconventional prepositions for an expression, as exemplified in (6):

(6) a. Jana

na sobotu z neděli
 on Saturday.ACC from Sunday.ACC

b. Standard Czech

ze soboty na neděli
 from Saturday.GEN to Sunday.ACC
 'from Saturday to Sunday'

This led to differing requirements for the case endings. The meanings of the prepositions do not mirror those that would be required by English syntax, meaning that this phrase cannot be attributed to the influence of English.

Adéla, Dana, and Jana, who are all members of the 1.5 or 2nd generation, were the only participants who used prepositions in an unconventional way. It is possible that this is attributable to intergenerational attrition/shift. Vašek (1996) attributes the interchange or omission of prepositions in American Czech to weakening awareness of their meanings. However, where participants do more clearly reflect English syntax, it is possible that the prepositions in question have either acquired meanings more compatible with those available in English or are simply used subconsciously to match the syntax of both languages.

4.1.3. The Nominative Becomes the Default

There is an increased frequency in the use of the nominative case in place of other syntactically required cases in diaspora Czech communities (Vašek 1996; Dutková 1998; Zajícová 2009). Zajícová (2009) regards this as attributable to a joint influence of language-contact-induced transfer and internal attrition processes. Spanish and Modern English do not have fully fledged case systems. Participants may forget or not know case endings due to lack of use and generational attrition/shift. With no similar system operating in English, as the syntax changes to more closely resemble English, use of case systems wanes and word order becomes more prominent as a feature (Larmouth 1974; Maher 1991; Zajícová 2009). Case endings tend to disappear throughout the generational attrition process amongst immigrant enclave communities (Maher 1991). Larmouth's (1974) study of immigrant Finnish speakers in Minnesota found that the case system is standard for first-generation speakers, optional in the second and third generation, and not consistently evident in the fourth. Schmid (2011) suggests that the tendency for complex case systems to merge and simplify over time becomes especially productive in situations of language contact. Yilmaz and Schmid (2019) discuss German and Croatian minority dialects in Italy, wherein young speakers have a tendency to over-generalize the case marker.²⁷

In this study, participants occasionally used unconventional case endings (Table 8). In South Australian Czech, only one 1st-generation participant used an unconventional case, with members of the 1.5 or 2nd generation producing the remainder of the unconventional case endings. This suggests that inter-generational attrition/shift may be playing a role here.

In the example in (7), the nouns *manžel* and *dcera* conventionally require an accusative case ending (*manžela* and *dceru*) but are instead in the nominative. As a 1.5-generation speaker, Adéla has instead used the resource of word order to derive meaning. It is possible that this is due to her knowledge of English, but it must be taken into consideration that this statement also exemplifies the unmarked word order in Czech.

(7) Adéla

mám manžel-Ø a dcera
have.1SG husband-NOM and daughter.NOM
'I have a husband and a daughter'

²⁷ Yilmaz and Schmid (2019: 198–209) attribute this attrition to language contact. Their explanation is cognitively based: they suggest that the additional language system leads to “demands of competition and limited cognitive resources”, which “can affect production, perception and comprehension”.

Table 8. Use of unconventional case endings

Participant	Number of uses
Adéla	6
Dana	1
Eva	2
Jana	1
Ivana	0
Kamila	0
Milada	0
Zuzana	0
Roman	0
Martin	0

In the following example, (8), the preposition *v* conventionally calls for a locative case ending (*Austrálii*). Adéla, however, uses a nominative case ending here (*Austrálie*).

- (8) Adéla
 jsme neměli rodinu v Austrálie
 AUX.1PL not.have.PL.PST family in Australia.NOM
 ‘we didn’t have family in Australia’

In example (9) below, Eva uses the accusative case (*rodinu*) where the preposition conventionally requires the dative case (*rodině*).

- (9) Eva
 kvůli rodinu
 because.of family.ACC
 ‘because of the family’

The following is an interesting occurrence, because Dana realizes that the preposition *s* ‘with’ requires an instrumental case ending on the noun *učitel* ‘teacher’ but does not apply this to *pan* ‘mister’, simply applying the nominative case in this scenario, (10a).

- (10) a. Dana
 s pan-Ø učitelem
 with mister-NOM teacher.INS
- b. Standard Czech
 s panem učitelem
 with mister.INS teacher.INS
 ‘with the teacher’

This tendency to eliminate oblique case inflection has also occurred in American Czech, where there is a tendency for the nominative- and accusative-case suffixes to be used where a different case is grammatically required (Henzl 1982: 42). In Dutková’s (1998: 632) study on the structural features of Texan Czech, she found that the “older generation” (pre-1945 group) “correctly” indicated case markings on translations in the Reduced Task²⁸ 78.7% of the time, with the “younger generation” (post-1945 group) indicating case markings “correctly” only 26.4% of the time.

4.1.4. Gender Distinction: Masculine and Feminine Are Swapped

In some European languages, and Arabic, there is a tendency for the masculine gender to be either overgeneralized, utilized in situations of unfamiliarity, or reanalyzed as a neutral form in the absence of overt morphological cues for a feminine classification, especially if the masculine is the default, unmarked form in that language (Dieser 2009; Brehmer and Rothweiler 2012; Albirini et al. 2013; Bianchi 2013; Cuza and Pérez-Tattam 2016). However, sometimes phonological cues for a feminine classification (i.e., an ending usually reserved for feminine forms, e.g., *-a*) result in unconventional use of the feminine agreement forms. The example below, (11), from Pereltsvaig’s (2004) paper on the absence of gender agreement in American Russian, shows this:

- (11) a. American Russian
 moj_a deduška
 my.F grandpa(M)
- b. Standard Russian
 moj-Ø deduška
 my-M grandpa(M)
 ‘my grandpa’
- (Pereltsvaig 2004: 90)

²⁸ The “reduced task” involved the translation of 20 sentences from English into Texan Czech and was aimed at eliciting features of “reduced” Czech (Dutková 1998).

Phrases which do not show gender concord occur in this dataset (see Table 9), but participants did not **necessarily** default to masculine gender use.

Table 9. Unconventional gender use (no gender concord)

Participant	Unconventional gender used
Adéla	1
Dana	0
Eva	0
Jana	1
Ivana	0
Kamila	0
Milada	2
Zuzana	2
Roman	1
Martin	0

It is interesting to compare this to similar situations, wherein use of phrases that do not show gender concord are presumably much more pronounced (Vašek 1996; Zajícová 2009, 2012). It may be the case that, as this community is much “younger” generationally, such a feature may not yet frequently occur.

The examples below demonstrate how the Czech South Australian community have used grammatical gender.

(12) Adéla

vím že moje brácha
 know.1SG that my.F brother.M.AN
 ‘I know that my brother’

It is likely that Adéla’s use of a feminine possessive pronoun can be attributed to a phonological cue for feminine classification from the noun. However, the noun is masculine animate.

(13) Jana

to jsou moje lidi
 it are my.PL.M.INAN/F/N people.NOM.M.AN
 ‘these are my people’

Here the masculine animate noun *lidi* ‘people’ (whose form is used in the spoken language—in the written language, it is *lidé* in the nominative) requires a possessive pronoun in the masculine animate plural. Jana instead uses the possessive for masculine inanimate, feminine, or neuter nouns. It is possible that the conventional ending was “forgotten” here due to attrition. It is also possible that Jana is using the accusative case here (wherein *moje lidi* would be a correct form for the masculine animate plural) rather than the required nominal case.

There are different forms for expressing ‘two’ in Czech, depending on the gender of the accompanying noun. The masculine form of ‘two’ is *dva*, and the feminine and neuter forms are represented by *dvě*. In example (14) below, Milada uses the feminine/neuter form rather than the masculine.

(14) Milada

mám	manžela	a	dvě	kluky
have.1SG	husband	and	two.F/N	boys.M

‘I have a husband and two boys’

4.1.5. Reflexive Pronouns: Disuse and Unconventional Use

In Czech, reflexive pronouns serve a variety of functions. They can derive a reflexive verb, a reciprocal verb, or a passive, impersonal, or intransitive verb from a transitive verb (Janda and Townsend 2000: 59). They can also represent a required component of a verb that only exists in accompaniment with *si* or *se* (a lexical reflexive) (Janda and Townsend 2000: 59).

Reflexive pronouns inflect for case; the dative case requires the reflexive pronoun form *si*, as in (15a), while the accusative case requires the reflexive pronoun form *se*, (15b).

(15) a. Dative case

Myju	si	ruce.
wash.1SG	REFL.DAT	hands.ACC

‘I wash my hands.’ (lit. ‘I wash for/to myself hands’)

b. Accusative case

Myju	se.
wash.1SG	REFL.ACC

‘I wash myself.’ (i.e., the entire self)

Unconventional reflexive pronouns can thus serve as an example of a loss of case distinction, especially in cases of transitive verbs that can be used re-

flexively. It is also possible that missing reflexive pronouns are more likely to occur with Czech lexical reflexives that are not reflexive in English, following English syntax and directly transferring the phrase over.

Adéla and Eva use the largest number of unconventional reflexive pronouns (see Table 10 on the following page). These participants are from the 1.5 and 2nd generation, and the other user of unconventional reflexive pronouns, Jana, is also from the 2nd generation. Thus, in this dataset the unconventional use of reflexive pronouns may be a result of intergenerational attrition/shift.

In one example, (16), Adéla uses the dative form of the reflexive pronoun with the verb *učit se* 'to learn' (lit. to teach oneself), for which the accusative form is required. It could be argued that the verb *učit* 'to teach' is transitive and, when used reflexively, represents 'to learn', maintaining the idea that utilization of an unconventional reflexive pronoun could represent loss of case distinction.

(16) Adéla

jsem	si	učila
AUX.1SG	REFL.DAT	learned

'I learned'

Example (17b) shows the way that the lexical reflexive verb *snažit se* 'to try' (in the sense of 'to strive') is conventionally used in Czech. Eva uses the verb without the reflexive pronoun, (17a). This verb does not require a reflexive in English, so it may be that grammatical replication is at play here.

(17) a. Eva

snažím	Ø	ted'ka
try.1SG	?	now

b. Standard Czech

snažím	se	ted'ka
try.1SG	REFL.ACC	now

'I'm trying now'

In the following example, (18), Jana uses the reflexive pronoun where it is not conventionally required.

(18) Jana

jak	se	může	říct
how	REFL.ACC	is.able	to.say

'how do I say this?'

Table 10. Forms of reflexive pronouns

Participant	Total REFL pronoun possibilities	Conventional REFL pronoun used	Unconventional REFL pronoun used		
			<i>se</i> vs. <i>si</i>	Not included	Used where not required
Adéla	8	5	1	2	0
Dana	28	28	0	0	0
Eva	15	8	1	5	0
Jana	19	17	0	0	2
Ivana	10	10	0	0	0
Kamila	23	23	0	0	0
Milada	10	10	0	0	0
Zuzana	13	13	0	0	0
Roman	7	7	0	0	0
Martin	6	6	0	0	0

Jana mixes the phrases *jak se říká/řekně* ‘how does one say’ and *jak můžu říct* ‘how can I say’ in a way that is not conventional in Czech.²⁹ It is possible that this is an example of redundancy of expression, a phenomenon that occurs when the speaker is not fully confident that the utterance will be parsed and decoded correctly and introduces more “instructional” elements to guide the hearer (Polinsky 1997: 398–99).

4.1.6. Syntax: English Influence?

Several participants adopted English construction types by choosing unconventional constructions and increasing the analytic nature of the sentence by utilizing verbs such as *jít* ‘to go’ and *dělat* ‘to do’ as auxiliaries. Sentences considered attestations are somewhat difficult to quantify here, as they represent a number of different phenomena—including use of an auxiliary + infinitive, in keeping with English syntax (see (19–20)), as well as unconventional word order (see (21)). Sentences produced would generally make sense to a Czech person, but they would not sound conventional.

Most participants who produced such attestations are in the 1.5 or 2nd generation (see Table 11 below).

Table 11. Non-Czech conventional word order/English word order/syntax

Participant	Attestations
Adéla	1
Dana	4
Eva	2
Jana	3
Ivana	0
Kamila	0
Milada	0
Zuzana	0
Roman	3
Martin	2

²⁹ This phrase is not necessarily grammatically incorrect, but it has a meaning different from what the speaker intended. The speaker intended to ask how to say a certain word in Czech, but the phrase produced can be used as a rhetorical question, e.g., ‘how can you say X is true?’

This could be evidence for their language development compared with those who arrived later as 1st-generation immigrants (Polinsky 2008: 334). It is possible that these people had divergent attainment of Czech as children, which represents intergenerational language attrition/shift (Huffines 1991; Burling 1992; Waas 1996; Polinsky 1997; Hickey 2010). Van Els (1986) posits that the main cause of language loss is not due to the individual forgetting elements of the language, but rather incomplete transfer between generations and thus incomplete acquisition (now called divergent attainment; cf. Kupisch and Rothman 2016; Polinsky 2018). Indeed, the former is a contributing cause to the latter.

Interestingly, one of the first-generation participants who produced such attestations mentioned that they had not been back to the Czech Republic for over seven years, possibly suggesting a lack of use of the language and thus some intragenerational attrition (Stoessel 2002; Clyne and Kipp 2006). This particular participant is also married to an individual with another non-English L1, which has a greater relative importance in terms of speaker population in Australia. It is possible that this other language is thus prioritized in terms of conversation together and with their children.

The example in (19) below shows how Eva utilizes Australian English syntactic structure and substitutes an Australian English word.

(19) Eva

on nechce **jít** **camping**
 he not.want to.go camping
 'he doesn't want to go camping'

In Australian English, in this context, one would not often say 'he doesn't want to camp', as such a phrasing has a perfective sense, but rather one would say 'he doesn't want to go camping', giving an imperfective sense to the phrase. In English, utilizing the second phrase gives a more accurate depiction of the activities involved in engaging in camping. The sentence becomes more analytic in utilizing the infinitive and a noun rather than simply using the verb.

This is then repeated by Dana in (20):

(20) Dana

tam můžeš **dělat i** **camping**
 there be.able.2SG to.do also camping
 'you can go camping there'

The unconventionality here focuses on the syntax, as the English word *camping* and its phonological alternations *kempink* and *kemping* do occur in Czech speech in the Czech Republic (see Appendix 8). Eva also utilizes the verb *kem-*

povat conventionally in the next sentence, perhaps in self-correction. However, after this, Dana continues to use the long form with the English vocabulary *dělat i camping* ‘to do camping’. This further shows English syntactic influence.

Increasingly analytic syntax is also evident in Texan Czech (Dutková 1998). Both generations in Dutková’s study found it difficult to produce the “correct” Standard Czech imperfective verb, with half of the older generation (pre-1945 group) and most of the younger generation (post-1945 group) opting for use of an auxiliary and an infinitive in its place, confirming Kučera’s (1989) observation of exactly this feature in American Czech (Dutková 1998: 64).

The syntax of Jana’s sentence in (21a) follows that of SVO English: ‘when (did) your kids go to school here?’ However, natural Czech speech requires a different word order: ‘when go your kids here to school?’ (21b). VSO and VOS sentences are the most natural word-order choices for Czech questions, with the WH-question word typically appearing at the beginning of the sentence (Janda and Townsend 2000). Syntactic change to further follow L2 sentence constructions and word order is also evident in Australian German (Waas 1996).

(21) a. Jana

kdy	vaše	děti	šli	do školky	tady
when	your	kids.NOM.F	went.PL.M.AN	to school	here

b. Standard Czech

Kdy	šly	vaše	děti	tady do školy?
when	went.PL.F	your	kids.NOM.F	here to school

‘When did your kids go to school here?’

In American Czech, sentence constructions and phrases often completely imitate those present in American English, and over time, a complete elimination of cases have led syntactic function to be derived from word order (Henzl 1982; Vašek 1996: 82). It would appear then that attrition processes play a role in the erosion of case endings, leading to a subsequent calquing of English word order, which then serves to further eliminate the need for the use of cases.

4.1.7. Tentative Article Formation

Czech has no distinctive article word class. In this data, participants use the demonstrative *ten* and the numeral *jeden* (and their derivatives) to form definite and indefinite articles, which is also attested in the Zajícová (2009) and Dutkova-Cope (2001a) data from Paraguay and Texas. This use of numerals and demonstratives to create a category non-existent in Czech may be an ex-

ample of filling a “grammatical gap”. The filling of grammatical gaps is posited as a reason for grammatical borrowing in situations of language contact, particularly among earlier scholars (Hale 1975; Heath 1978; Hill and Hill 1981; Campbell 1993). The numeral *jeden* ‘one’ is utilized as an indefinite article in American Czech (Vašek 1996: 81).

It is mostly Kamila, Zuzana, and Martin who produce a possible tentative article (see Table 12 below).

Table 12. Tentative article formation

Participant	Unconventional use of demonstrative as article
Adéla	0
Dana	1
Eva	0
Jana	0
Ivana	0
Kamila	3
Milada	0
Zuzana	2
Roman	1
Martin	2

Two examples found in the data for this study are shown in (22) and (23):

(22) Zuzana

To byly takové ty koule,
it were such/some.sort DEM.NOM.PL.F ball.NOM.PL.F
to jsou ty české
it are DEM.NOM.PL.F Czech.NOM.PL.F
‘It was some sort of balls, some sort of Czech’

(23) Kamila

mám ty vnoučata
have.1SG DEM.ACC.PL grandchild.ACC.PL.N
‘I have the grandchildren’

However, it is also possible that the interlocutors are speaking Common Czech, wherein *ten* and its derivatives are used as definite articles or pronouns (Janda and Townsend 2000). This use of the demonstrative as a definite article is a tendency which is increasingly occurring (Zíková 2017). Zíková (2017) posits that persistence of the referent (speaker-evaluated local importance of the referent in the narrative) may be an explanatory factor in this grammaticalization process. It is possible that the grammaticalization process in the Czech Republic is accelerated by the influence of English as a language of international prestige. Other lexical and grammatical “Anglicisms” have been borrowed into the Czech language in the past, including calquing of idioms (*být in [být modní]* ‘to be in fashion’), direct lexical borrowing (*billboard, newsroom*), and modifications of syntactic patterns (*ten pohled je prostě dech beroucí* ‘the view is simply breathtaking’³⁰) (Bozděchová 1997: 276–77; Tarnýíková 2009: 205; see also Warmbrunn 1994; Gester 2001; Markova 2018).

Cvrček (2015: 174–75) states that *ten* and *jeden* can be used as determiners to express specificity, or alternatively, express distance from the referent. It is also a possibility that uses of the demonstrative and numerals in this dataset express these concepts. In addition, the use of this feature by primarily Kamila, Martin, and Zuzana, all first-generation participants, means that it is unlikely that this feature is an example of intergenerational shift.

4.1.8. Summary of Data and Diaspora Comparison

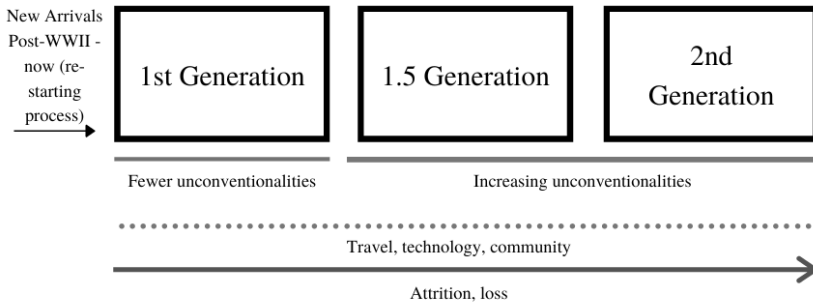
Each of the grammatical features found have been separately discussed and compared with other diasporic communities, and qualitative judgements have been made on the origins of each feature based on both the specific examples shown and the data frequencies.

It would appear that many of the grammatical unconventionalities occurring in South Australian Czech represent grammatical replication rather than borrowing as defined in §1; the way grammatical information is conveyed syntactically is altered rather than morphemes being directly borrowed. This is similar to Zajícová’s (2012) observation about Paraguayan Czech; Spanish has had more syntactic rather than morphological influence on Czech. It is likely that at least some of the features observed are attributable to transfer induced by language contact, while other features are explained by attrition processes, especially through incomplete intergenerational acquisition. Some features attributed to language contact or attrition processes by previous authors are possibly due to use of Common Czech.

When considering the similarity of the grammatical unconventionalities occurring in the diaspora communities (see Table 1 on p. 7), it is important

³⁰ Tarnýíková (2009) discusses how a Czech sentence would typically use a V [lex] predication here, rather than the stative BE-predication typical of English.

South Australian Czech



American, Paraguayan Czech

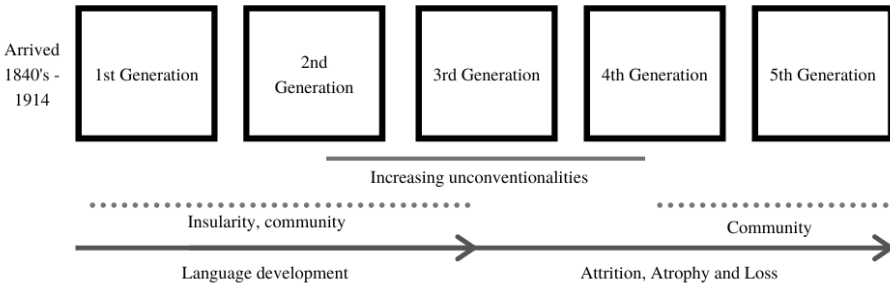


Figure 2. South Australian Czech and American and Paraguayan Czech language (adapted from Castle 2021)

to recognize the effects of a difference in time period of migration. Figure 2 above displays a summary of the current situation in South Australian Czech in comparison with American and Paraguayan Czech language situations. The modern South Australian Czech community consists of primarily 1st–2nd generation adults, who have immigrated between WWII and now. Grammatical unconventionalities tend to increase in the 1.5 and 2nd generation, though travel, technology, and community provide opportunities for language use and therefore a degree of language maintenance. However, as time passes and generations continue, attrition and loss occur (Castle 2021).³¹

³¹ It is important to note that, though attrition and transfer are occurring here, maintenance activities still allow for a greater degree of language maintenance than if they were not engaged in at all (Herdina and Jessner 2002; Jessner 2003). It is also important to note that, while maintenance-assistive technologies including travel and phones/ the internet are available to speakers now, many of the older participants were completely cut off from communication with relatives and friends in the Czech Republic during Communist Party rule.

The American and Paraguayan communities studied by Henzl (1982), Vašek (1996), Dutková (1998), and Zajíčová (2009, 2012), on the other hand, are much further along in the language attrition process. Czechs in these communities arrived between the 1840s and 1914, and therefore their adult descendants are now in the third, fourth, and fifth generations (Dutková 1998; Zajíčová 2009). The language was maintained more strongly amongst the first and second generations as the communities were insular at the time, particularly in the Texas Czech community, due to factors including the establishment of community professional, social, and religious institutions; reinforcement of ethnic identity regarding language use; adherence to traditions and language planning; the prevalence of endogamous marriages; maintenance of contact with the homeland through letters from the Czech and Moravian lands; and an ideology of *národnost*³² (Eckert and Hannan 2009: 103, 133). However, forces for assimilation during and after WWII created a distancing of the second and third generations from their language, which was a factor in attrition and atrophy (Eckert 2006; Eckert and Hannan 2009; Vaculík 2009). In the current era, many Czech descendants seek to experience community together, though the language is mostly lost apart from some key greetings and phrases (Hannan 2004; Cope 2011; see Castle 2021 for more details on this).

4.2. Analysis: Contact-Induced Replication or Attrition?

In this section, the qualitative conclusions reached about language-contact-induced borrowing are further considered by utilizing steps to establish that contact-induced structural change has occurred (Thomason 2001: 93–94). These steps to establish **structural** change, or **replication**, are able to be used as it is replication rather than borrowing that has occurred here (§4.1.8). It is made clear here that the steps are adapted to identify the source of potential unconventionalities—the focus is on whether these features are contact-induced rather than representing community-wide change (see §1).³³

The paraphrased steps/rules are as follows:

1. Cases for contact-induced structural changes must be supported by other instances of structural interference from the same source language in the same receiving language: there must be more than one type of case.

³² This is described by Eckert and Hannan (2009: 103) as a vision that was focused on the “Czech language of national literature”.

³³ Of course, such features may represent community-wide change, but proving such a change is outside the scope of this article.

2. The source and receiving languages must be shown to be in intimate enough contact to make structural interference possible.
3. Structural features shared by the proposed source and receiving languages need to be identified.
4. Prove that the proposed interference features were not present in the receiving language before coming into contact with the source language.
5. Prove that the proposed interference features were present in the source language before coming into contact with the receiving language.
6. Consider plausible internal motivations for the changes and the “very real possibility of multiple causation”.

(Thomason 2001: 93–94)

In terms of step 1, there are several types of potential cases that have been identified (§4.1). Participants’ languages are in intimate contact and have been for several generations (step 2). They utilize both the source and recipient languages in their daily lives, with the source language being used by the wider society and recipient language in their homes, with family and friends, and at the Club (Table 4 on p. 11). The relevant structural features of the two languages are presented in Table 13³⁴ (step 3). Table 13 can also be utilized to position each proposed change with respect to the host linguistic system and detect presumed causes, as well as showing whether the proposed interference features were not present in the pre-contact variety and present in the source variety prior to contact (steps 4 and 5).

In the discussion below, I analyze and explain each feature, with consideration of internal motivations (step 6). The overt subject feature is not present in Czech, as Slavic languages are pro-drop (Haspelmath et al. 2001). However, it does occur in Common Czech. Overt subject marking is required in English (Haspelmath et al. 2001). This feature could be contact-induced, as well as a result of attrition, but it is also possible that it represents use of Common Czech.

Slavic languages tend to have fully fledged case systems, whereas case inflection in English is present only in some pronouns. It is possible that the unconventionalities observed in the dataset are a result of grammatical replication of syntax due to attrition of case endings. The participants who produced unconventional case endings were in the 1.5 and 2nd generation (§4.1.3). This interacts with an increasingly analytic syntax; the roles of core syntactic cases

³⁴ The grey bars for prepositional system and reflexive pronoun phenomena are included because these features could not be analyzed in the same way (the unconventionalities observed represent several phenomena within these categories, so a Yes/No/Maybe answer was not possible here).

Table 13. Presence of feature in Czech and English with possible causes

Feature amongst South Australian Czech participants	Is feature present in Common Czech?	Is feature present in Australian English?	Is feature contact-induced?	Is feature a result of attrition/shift?
Use of overt subject	Yes	Yes	Maybe	Maybe
Prepositional system				
No case system	No	Mostly	No	Yes
No gender distinction	No	Yes	No	Yes
Reflexive pronoun phenomena				
Analytic syntax	No	Yes	Yes	Maybe
Articles required	No	Yes	Maybe	Maybe

become increasingly redundant in speech with a rigid word order to provide grammatical information. It is not possible here to establish the directionality: whether the language has become more analytic in response to divergent attainment (Andersen 1982; §4.1.1), or whether the case system is rendered redundant with a fixed word-order system providing the grammatical information.

Czech has three grammatical genders and an animacy distinction. English does not have a productive gender system (excepting some nouns and pronouns). It is possible that the observed unconventional use of grammatical gender represents attrition processes and language contact, as in Zajíčová 2009, where a frequent use of the nominative in place of other cases is attributed to a “combined influence of Spanish and attrition” (p. 144). However, the extremely small number of attestations could suggest that the community is still quite young in comparison with other Czech diaspora communities in terms of generation and therefore aspects of intergenerational attrition.

The reflexive pronoun could not be analyzed in the same way, because the unconventionalities represent three phenomena: use when not conventional, non-use when conventional, and use of *se* or *si* (§4.1.5). There is some evidence here for attrition processes, as all attestations of these unconventionalities are from the 1.5 and 2nd generations. The prepositional system also could not be analyzed in this way because the unconventionalities represent several phenomena: inclusion where unconventional, non-inclusion where unconventional, and unconventional choice. All attestations of these unconventionalities also come from the 1.5 and 2nd generation, providing evidence for the role of attrition.

Articles are not required in Czech (Dryer 2013). However, in Common Czech, demonstratives are used more often in places where there would be articles in other languages (Janda and Townsend 2000). This is part of a process of grammaticalization into articles (see §4.1.7). Articles are required in English (Dryer 2013). It is thus possible that article use could represent attrition, grammatical replication, or use of Common Czech.

Table 13 shows that several of the proposed changes were not present in the pre-contact variety, including the lack of a case system, lack of gender distinction, analytic syntax, and the requirement of articles. Overt subject use and the extended use of demonstratives are possible in Common Czech and may thus represent internal variation (step 6). However, it remains possible that this also represents contact-induced grammatical replication (see §4.1.1, 4.1.7). These features were all possible in the source language prior to contact between the South Australian Czech community and Australian English.

Divergent attainment is particularly likely to be a contributor to the instability of prepositions, loss of case distinction, loss of gender distinction, and increased analytic nature of the language, as the speakers engaging in these were primarily from the 1.5 and 2nd generation. Widely recognized signs of a

language undergoing attrition include increased analytic nature no matter the source language structure, issues with loss of case distinction and increase in the use of the nominative case, preposition instability, and loss of gender distinction (Andersen 1982; Polinsky 1997; Zájíková 2009). However, it is also likely that the speech of divergent attainers is influenced by their dominant language. Indeed, some authors consider this to be part of the attrition process (Sharwood Smith and Kellerman 1986; Grosjean and Py 1991; Pavlenko 2000; Gürel 2002; Schmid and Keijzer 2009; Cherciov 2013).

The Czech South Australian community is moving through processes of language shift. The Czech South Australian community is at the attrition stage, though the possibility of an influx of new community members from the Czech Republic keeps the cycle continuing (Castle 2021; see also Figure 2, this paper).

Excepting those possibly created by internal motivations, all of the unconventionalities discovered are the product of the sociolinguistic situation induced by language contact. A key premise of Dynamic Systems Theory is applicable here: a dynamic system is a set of variables that mutually affect each other's changes over time (van Geert 1994; Herdina and Jessner 2002). In this case, contact-induced transfer and attrition represent those variables; they have a somewhat symbiotic relationship, influencing one another and acting jointly to produce the features observed. Attrition occurs in the contact situation due to the introduction and required use of the majority language, and thus ever-decreasing frequency of use of one's own language, possibly resulting in language loss and language death. As resources from one language are lost due to attrition, resources from the other language are borrowed. For example, as the resource of a full-fledged case system is lost in Czech, there is a tendency to use a more rigid word order, which is a feature of English but is also a tendency of languages undergoing attrition and shift.

5. Conclusion

Observation session data on individuals in the Czech South Australian community was collected and analyzed to detect whether contact-induced **borrowing** and **grammatical replication** innovations occurred. Participants displayed several grammatical features in their speech, including increasing the analytic nature of the language, use of the overt subject, loss of gender distinctions, preposition instability, tentative article formation, and loss of case distinctions. These features match those that have occurred in America and Paraguay (Henzl 1982; Vašek 1996; Dutková 1998; Zájíková 2009; 2012). Grammatical replication rather than borrowing (Heine and Kuteva 2008; Kuteva 2017) has occurred in South Australian Czech, similar to Zájíková's (2012) study.

Despite similar findings as those in other diaspora communities, this paper notably analyzes a different period of migration and thus examines a language contact situation in the era of increased connectivity in terms of travel and the availability of phones and internet access (Keijzer 2020). It might be predicted that interconnectivity would mitigate against language attrition and contact-induced transfer, but despite this, the study demonstrates that unconventionalities are occurring at the level of morphology and syntax. However, it is also noted that engagement in maintenance activities does mitigate against attrition and transfer more than if such activities are not engaged in (Herdina and Jessner 2002; Jessner 2003). By adding data from a vastly different temporal and geographical context, this study aids in developing a more nuanced understanding of how and why speakers use different resources from between their languages.

Through analysis using Thomason's (2001) steps to identify instances of contact-induced structural change and dynamic systems theory, it is posited that at least increasingly analytic syntax, overt subject usage, and tentative article formation are partially attributable to language contact and grammatical replication. This paper therefore adds to the literature which states that it is possible for language-contact-induced grammatical borrowing to occur, while also positing that contact-induced language transfer and shift and attrition processes exist in a symbiotic relationship.

Future research could involve an analysis of whether innovations have resulted in community-wide propagations. This would require a larger sample size, more time analyzed per speaker, and a large Czech-habitant comparison group to allow researchers to be able to make generalizations and stronger assertions about causation. Other future research could include the study of Czech in contact with a language with equal or richer morphology.

Sources

- Australian Bureau of Statistics. (2017) "2016 census Quickstats country of birth: People in South Australia who were born in Czech Republic". Available at: https://www.abs.gov.au/census/find-census-data/quickstats/2016/3302_4. Last accessed 7 July 2022.
- . (2022) TableBuilder. Available at: <https://www.abs.gov.au/census/guide-census-data/about-census-tools/tablebuilder>. Last accessed 7 July 2022.
- Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kovářiková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondříčka, and Adrian Zasina. (2018) "Korpus SYN, verze 7 z 29.11.2018". Prague: Ústav Českého národního korpusu FF UK. Available at: <http://www.korpus.cz/>. Last accessed 20 July 2022.

- Křen, Michal, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Koček, Dominika Kovářiková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal. (2020) "SYN2020: reprezentativní korpus psané češtiny". Prague: Ústav Českého národního korpusu FF UK. Available at: <http://www.korpus.cz/>. Last accessed 20 July 2022.
- Machálek, Tomáš. (2014). KonText – aplikace pro práci s jazykovými korpusy [KonText – An application for working with language corpora]. FF UK, Prague. Available at: <http://kontext.korpus.cz>. Last accessed 20 July 2022.
- . (2019). Slovo v kostce – agregátor slovních profilů [Word at a Glance – Aggregator of word profiles]. FF UK, Prague. Available at: <http://korpus.cz/slovo-v-kostce/>. Last accessed 20 July 2022.

References

- Albirini, Abdulkafi, Elabbas Benmamoun, and Brahim Chakrani. (2013) "Gender and number agreement in the oral production of Arabic heritage speakers". *Bilingualism* 16(1): 1–18.
- Altenberg, Evelyn P. (2010) "Assessing first language vulnerability to attrition". Herbert W. Seliger and Robert M. Vago, eds. *First language attrition*. Cambridge: Cambridge University Press, 189–206.
- Andersen, Roger W. (1982) "Determining the linguistic attributes of language attrition". Richard D. Lambert and Barbara F. Freed, eds. *The loss of language skills*. Rowley, MA: Newbury House Publishers, 83–118.
- Auty, Robert. (1976) "Problems of the formation and development of the Czech Literary Language". Thomas Magner, ed. *Slavic linguistics and language teaching*. Columbus, OH: Slavica Publishers, 82–88.
- Bermel, Neil. (2000) *Register variation and language standards in Czech*. Munich: Lincom Europa. [LINCOM Studies in Slavic Linguistics, 13.]
- Bělič, Jaromír. (1959) "Bojujme za upevňování a šíření hovorové češtiny" [We fight for the consolidation and dissemination of colloquial Czech]. *Český jazyk a literatura* 9: 433–41.
- . (1960) "Patnáct let nové republiky a český jazyk" [15 years of the new Republic and the Czech language]. *Naše řeč* 43(5–6): 129–34.
- Bianchi, Giulia. (2013) "Gender in Italian-German bilinguals: A comparison with German L2 learners of Italian". *Bilingualism* 16(3): 538–57.
- Bozděchová, Ivana. (1997) "Vliv angličtiny na češtinu" [The influence of English on Czech]. František Daneš, Jarmila Bachmannová, and Světlá Čmejř, eds. *Český jazyk na přelomu tisíciletí*. Prague: Academia, 271–79.
- Brehmer, Bernhard and Monika Rothweiler. (2012) "The acquisition of gender agreement marking in Polish: A study of bilingual Polish-German-speaking children". Kurt Braunmüller and Christoph Gabriel, eds. *Mul-*

- tilingual individuals and multilingual societies*. Amsterdam: John Benjamins Publishing Company, 81–100.
- Brouček, Stanislav, Ivo Barteček, Veronika Beranská, Tomáš Grulich, Marek Jakoubek, Jana Kočí, Petr Lozoviuk, Jaroslav Marek-Vejvoda, Lubomír Martinek, Andrej Sulitka, Zdeněk Uherek, and Ota Ulč. (2019) *Česká Republika a diaspora: Co bylo a co bude?* [The Czech Republic and the diaspora: What was and what will be?]. Prague: Etnologický ústav AV ČR.
- Burling, Robbins. (1992) *Patterns of language: Structure, variation, and change*. San Diego, CA: Academic Press, Inc.
- Campbell, Lyle. (1993) "On proposed universals of grammatical borrowing". *Historical Linguistics 1989: Papers from the 9th International Conference on Historical Linguistics*. Amsterdam: John Benjamins Publishing Company, 91–109. [Current Issues in Linguistic Theory, 106.]
- Castle, Chloe. (2021) "Language loyalty and language purity in a language contact situation: South Australian Czech". *Journal of Slavic linguistics* 29(1): 1–44.
- Charles Sturt Council. (2019) "Czechoslovak Club in SA Inc". Available at: https://sacommunity.org/org/201434-Czechoslovak_Club_in_SA_Inc. Last accessed 5 July 2022.
- Cherciov, Mirela. (2013) "Investigating the impact of attitude on first language attrition and second language acquisition from a Dynamic Systems Theory perspective". *International journal of bilingualism* 17(6): 716–33.
- Clyne, Michael. (1982) *Multilingual Australia*. Melbourne: River Seine.
- . (2003) *Dynamics of language contact*. Cambridge: Cambridge University Press.
- Clyne, Michael and Sandra Kipp. (1996) "Language maintenance and language shift in Australia, 1991". *Australian review of applied linguistics* 19(1): 1–19.
- . (2006) "Australia's community languages". *International journal of the sociology of language* 180: 7–21.
- Cope, Lida. (2011) "From ethnocultural pride to promoting the Texas Czech Vernacular: Current maintenance efforts and unexplored possibilities". *Language and education* 25(4): 361–83.
- Cuza, Alejandro and Rocío Pérez-Tattam. (2016) "Grammatical gender selection and phrasal word order in child heritage Spanish: A feature re-assembly approach". *Bilingualism* 19(1): 50–68.
- Cvrček, Václav. (2015) "Morfologie" [Morphology]. Václav Cvrček and a collective of authors, eds. *Mluvnice současné češtiny: Jak se píše a jak se mluví* [A grammar of contemporary Czech: How to write and how to speak]. Prague: Univerzita Karlova v Praze Nakladatelství Karolinum, 153–353.
- Dieser, Elena. (2009) *Genuserwerb im Russischen und Deutschen: Korpusgestützte studie zu ein- und zweisprachigen kindern und erwachsenen* [Genus acquisition

- in Russian and German: Corpus-based study on bilingual children and adults]. Munich: Otto Sagner.
- Doğruöz, A. Seza and Ad Backus. (2009) "Innovative constructions in Dutch Turkish: An assessment of ongoing contact-induced change". *Bilingualism: Language and cognition* 12(1): 41–63.
- Dryer, Matthew S. (2013) "Definite articles". Matthew S. Dryer and Martin Haspelmath, eds. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: <http://wals.info/chapter/37>.
- Dutková, Ludmila. (1998) *Texas Czech: An ethnolinguistic study*. Ph.D. dissertation, University of Arizona.
- Dutkova-Cope, Lida. (2001a) "The language of Czech Moravians in Texas: Do you know what *párknu káru u hauza* means?" *Southwest journal of linguistics* 20(2): 51–84.
- . (2001b) "Texas Czech: The language of Texans who say they speak 'a different type of Czech'". *Southwest journal of linguistics* 20(1): 29–64.
- Eckert, Eva. (2006) *Stones on the prairie: Acculturation in America*. Bloomington, IN: Slavica Publishers.
- Eckert, Eva and Kevin Hannan. (2009) "Vernacular writing and sociolinguistic change in the Texas Czech community". Mark Richard Lauersdorf and Curt Woolhiser, eds. "North American contributions to Slavic sociolinguistics". Special double issue, *Journal of Slavic linguistics* 17(1/2): 87–161.
- Eckertová, Eva. (2017) "Čeština v Americe" [Czech in America]. Petr Karlík, Marek Nekula, and Jana Pleskalová, eds. *CzechEncy: Nový encyklopedický slovník češtiny* [CzechEncy: New encyclopedic dictionary of Czech]. Brno: Centrum zpracování přirozeného jazyka. Available at: <https://www.czechency.org/slovník/%C4%8CE%C5%A0TINA%20V%20AMERICE>. Last accessed 1 December 2021.
- Embassy of the Czech Republic in Canberra. (2021) "Czech community in Australia and New Zealand". Available at: https://www.mzv.cz/canberra/en/czech_presence/index.html. Last accessed 5 July 2022.
- Fischer, Olga. (2004) "What counts as evidence in historical linguistics?". *Studies in language* 28(3): 710–40.
- Gester, Silke. (2001) *Anglizismen im Tschechischen und im Deutschen: Bestandsaufnahme und empirische Analyse im Jahr 2000* [Anglicisms in Czech and German: Inventory and empirical analysis in 2000]. Frankfurt am Main: Lang.
- Grosjean, François and Bernard Py. (1991) "La restructuration d'une première langue: l'intégration de variantes de contact dans la compétence de migrants bilingues" [The restructuring of a first language: the integration of contact variants in the competence of bilingual migrants]. *La linguistique* 27(2): 35–60.

- Gürel, Ayşe. (2002) *Linguistic characteristics of second language acquisition and first language attrition: Turkish overt versus null pronouns*. Ph.D. dissertation, McGill University.
- Hale, Kenneth. (1975) "Gaps in grammar and culture". Dale M. Kinkade, Kenneth L. Hale, and Oswald Werner, eds. *Linguistics and anthropology: In honor of C. F. Voegelin*. Lisse: The Peter de Ridder Press, 295–315.
- Hannan, Kevin. (2004) "From one monolingualism to another: Ethnic assimilation as a product of language displacement in Czech-Moravian Texas". *Český lid* 91(3): 235–52.
- Haspelmath, Martin, Ekkehard König, Wulf Oesterreicher, and Wolfgang Rast, eds. (2001) *Language typology and language universals: An international handbook*. Vol. 2. Berlin: Walter de Gruyter.
- Heath, Jeffrey. (1978) *Linguistic diffusion in Arnhem Land*. Canberra: Australian Institute of Aboriginal Studies.
- Heine, Bernd and Tania Kuteva. (2003) "On contact-induced grammaticalisation". *Studies in language* 27(3): 529–72.
- . (2005) *Language contact and grammatical change*. Cambridge: Cambridge University Press.
- . (2008) "Constraints on contact-induced linguistic change". *Journal of language contact* 2(1): 1–34.
- . (2010) "Contact and grammaticalization". Raymond Hickey, ed. *The handbook of language contact*. West Sussex: Wiley-Blackwell, 86–105.
- Henzl, Vera. (1982) "American Czech: A comparative study of linguistic modification in immigrant and young children speech". Roland Sussex, ed. *The Slavic languages in emigre communities*. Carbondale, IL: Linguistic Research, 33–46.
- Herdina, Philip and Ulrike Jessner. (2002) *A dynamic model of multilingualism: Perspectives of change in psycholinguistics*. Clevedon: Multilingual Matters.
- Hickey, Raymond. (2010) "An assessment of language contact in the development of Irish English". Jacek Fisiak, ed. *Linguistic change under contact conditions*. Berlin: Mouton De Gruyter, 109–30.
- Hill, Jane H. and Kenneth C. Hill. (1981) "Variation in relative clause construction in Modern Nahuatl". Frances Karttunen, ed. *Nahuatl studies in memory of Fernando Horcasitas* [Texas Linguistics Forum 18]. Austin: University of Texas, 89–104.
- Hlavac, Jim. (2000) *Croatian in Melbourne: Lexicon, switching, and morphosyntactic features in the speech of second-generation bilinguals*. Ph.D. dissertation, Monash University.
- Ho-Dac, Tuc. (1996) *Languages in contact: Vietnamese-English code-switching in Melbourne*. Ph.D. dissertation, Monash University.
- . (2003) *Vietnamese-English bilingualism: Patterns of code-switching*. London: Routledge Curzon.

- Huffines, M. (1991) "Pennsylvania German: Convergence and change as strategies of discourse". Herbert W. Seliger and Robert M. Vago, eds. *First language attrition*. Cambridge: Cambridge University Press, 125–38.
- Janda, Laura A. and Charles E. Townsend. (2000) *Czech*. Munich: Lincom Europa.
- Jessner, Ulrike. (2003) "A dynamic approach to language attrition in multilingual systems". Vivian Cook, ed. *Effects of the second language on the first*. Clevedon, UK: Channel View Publications, 234–46.
- Keijzer, Merel. (2020) "First language attrition in the twenty-first century". Evangelia Adamou and Yaron Matras, eds. *The Routledge Handbook of Language Contact*. London: Routledge, 221–33.
- Kopečný, František. (1949) "Spisovný jazyk a jeho forma hovorová" [Literary language and its colloquial form]. *Naše řeč* 33(1–2): 14–22.
- Kupisch, Tanja and Jason Rothman. (2016) "Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers". *International journal of bilingualism* 22(5): 564–82. DOI 10.1177/1367006916654355
- Kuteva, Tania. (2017) "Contact and borrowing". Adam Ledgeway and Ian Roberts, eds. *The Cambridge handbook of historical syntax*. Cambridge: Cambridge University Press, 163–86.
- Kučera, Karel. (1989) *Český jazyk v USA* [The Czech language in the USA]. Prague: Charles University.
- Labov, William. (1972) *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- . (1990) "The intersection of sex and social class in the course of linguistic change". *Language variation and change* 2(2): 205–54.
- Larmouth, Donald W. (1974) "Differential interference in American Finnish cases". *Language* 50(2): 356–66.
- Maher, Julianne. (1991) "A crosslinguistic study of language contact and language attrition". Herbert W. Seliger and Robert M. Vago, eds. *First language attrition*. Cambridge: Cambridge University Press, 67–84.
- Markova, Elena M. (2018) "Assimilation of borrowings in Slavic languages (in the aspect of Russian-Czech-Slovak comparison)". *RUDN Journal of language studies semiotics and semantics* 9(4): 896–903.
- Matras, Yaron and Jeanette Sakel. (2007) "Investigating the mechanisms of pattern replication in language convergence". *Studies in language* 31(4): 829–65.
- Migration Heritage Centre. (2010) "Australia's migration history". Available at: <https://www.migrationheritage.nsw.gov.au/belongings-home/about-belongings/australias-migration-history/index.html>. Last accessed 3 July 2022.
- Migration Museum. (2020) "Czechs in South Australia". Available at: <http://adelaide.sa.gov.au/subjects/czechs-in-south-australia>. Last accessed 3 July 2022.

- Milroy, L. (1987) *Observing and analysing natural language*. Cambridge: Basic Blackwell.
- National Museum Australia. (2021) "White Australia Policy". Available at: <https://www.nma.gov.au/defining-moments/resources/white-australia-policy>. Last accessed 3 July 2022.
- Pauwels, Anne. (1988) "The future of ethnic languages in Australia". *International journal of the sociology of language* 72: 5–14.
- Pavlenko, Aneta. (2000) "L2 influence on L1 in late bilingualism". *Issues in applied linguistics* 11(2): 175–205.
- Pereltsvaig, Asya. (2004) "Agreement in the absence of agreement: Gender agreement in American Russian". Danijela Stojanović, ed. "Psycholinguistics in Slavic". Special issue, *Cahiers linguistiques d'Ottawa* 32: 87–107.
- Pintová, Pavlína. (2009) *Czechs in Texas: A historical analysis of Czech language maintenance among the early immigrants and their descendants*. B.A. thesis, Masaryk University.
- Polinsky, Maria. (1997) "American Russian: Language loss meets language acquisition". Wayles Browne, Ewa Dornisch, Natasha Kondrashova, and Draga Zec, eds. *Annual workshop on formal approaches to Slavic linguistics: The Cornell meeting 1995*. Ann Arbor: Michigan Slavic Publications, 370–406.
- . (2008) "Relative clauses in heritage Russian: Fossilisation or divergent grammar?" Andrei Antonenko, John F. Bailyn, and Christina Y. Bethin, eds. *[Formal] Approaches to [Slavic] linguistics: The Stony Brook meeting 2007*. Ann Arbor: Michigan Slavic Publications, 333–58.
- . (2018) *Heritage languages and their speakers*. Cambridge: Cambridge University Press.
- Preston, Dennis. (1982) "How to lose a language". *Interlanguage studies bulletin* 6(2): 64–87.
- Rakusan, Jaromira. (1993) "Code mixing as a vehicle of register: A case of 'Old Chicago Czech'". *Canadian Slavonic papers/Revue canadienne des slavistes* 35(3/4): 275–90.
- Romaine, Suzanne. (2005) "Language-contact studies". Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier, and Peter Trudgill, eds. *Sociolinguistics/Soziolinguistik. An international handbook of the science of language and society*. Berlin: de Gruyter, 49–58.
- Rumbaut, Rubén. G. and Kenji Ima. (1988) "The adaptation of Southeast Asian refugee youth: A comparative study. Final report to the Office of Resettlement". U.S. Department of Health and Human Services, Family Support Administration, Washington, D.C. [ERIC Document Service Reproduction Service, no. ED 299 372.]
- Rumbaut, Rubén. (1994) "The crucible within: Ethnic identity, self-esteem, and segmented assimilation among children of immigrants". *International migration review* 28(4): 748–94.

- Rumbaut, Rubén. (1997) "Paradoxes (and orthodoxies) of assimilation". *Sociological perspectives* 40(3): 481–511.
- Rumbaut, Rubén. (2004). "Ages, life stages, and generational cohorts: Decomposing the immigrant first and second generations in the United States." *International migration review* 38(3): 1160–1205.
- Schmid, Monika. (2011) *Language attrition*. Cambridge: Cambridge University Press.
- Schmid, Monika and Merel Keijzer. (2009) "First language attrition and reversion among older migrants". *International journal of the sociology of language* 200: 83–101.
- Seliger, Herbert W. and Robert M. Vago. (1991) "The study of first language attrition". Herbert W. Seliger and Robert M. Vago, eds. *First language attrition*. Cambridge: Cambridge University Press, 3–16.
- Sharwood Smith, Mike. (1989) "Crosslinguistic influence in language loss". Kenneth Hyltenstam and Loraine K. Obler, eds. *Bilingualism across the life-span*. Cambridge: Cambridge University Press, 185–201.
- Sharwood Smith, Mike and Eric Kellerman. (1986) *Crosslinguistic influence in second language acquisition*. New York: Pergamon Press.
- Sharwood Smith, Mike and Paul van Buren. (1991) "First language attrition and the parameter setting model". Herbert W. Seliger and Robert M. Vago, eds. *First language attrition*. Cambridge: Cambridge University Press, 17–30.
- Stoessel, Saskia. (2002) "Investigating the role of social networks in language maintenance and shift". *International journal of the sociology of language* 153: 93–131.
- Tarnýřková, Jarmila. (2009) "English borrowings in Czech: Health to our mouths?". *Brno studies in English* 35(2): 199–213.
- Thomason, Sarah G. (2001) *Language contact: An introduction*. Washington, D.C.: Georgetown University Press.
- . (2014) "Contact-induced language change and typological congruence". Juliane Besters-Dilger, Cynthia Dermarkar, Stefan Pfänder, and Achim Rabus, eds. *Congruence in contact-induced language change: Language families, typological resemblance, and perceived similarity*. Berlin: De Gruyter, 201–18.
- Torres Cacoullós, Rena and Catherine E. Travis. (2018) *Bilingualism in the community: Code-switching and grammars in contact*. Cambridge: Cambridge University Press.
- Vaculík, Jaroslav. (2009) *České menšiny v Evropě a ve světě* [Czech minorities in Europe and the world]. Prague: Libri.
- van Els, Theo. (1986) "An overview of European research on language attrition". Theo van Els, Kees De Bot, and Bert Weltens, eds. *Language attrition in progress*. Dordrecht: Foris Publications, 1–18.
- van Geert, Paul. (1994) *Dynamic systems of development: Change between complexity and chaos*. New York, NY: Harvester Wheatsheaf.

- Vašek, Antonín. (1996) "On language acculturation in American Czechs". *Sborník prací Filozofické fakulty Brněnské univerzity* 22(1): 71–87.
- Waas, Margit. (1996) *Language attrition downunder*. Berlin: Peter Lang.
- Warmbrunn, Jürgen. (1994) *Englische lexikalische Entlehnungen im Wortschatz der tschechischen Gegenwartssprache* [English lexical borrowings in the vocabulary of the contemporary Czech language]. Münster/Hamburg: Lit Verlag.
- Wei, Li. (2013) "Conceptual and methodological issues in bilingualism and multilingualism research". Tej K. Bhatia and William C. Richie, eds. *The handbook of bilingualism and multilingualism*. 2nd ed. Malden, MA: Wiley-Blackwell, 26–51.
- Weinreich, Uriel. (1953/1968) *Languages in contact: Findings and problems*. 9th ed. The Hague: Mouton. [Publications of the Linguistic Circle of New York, no. 1.]
- Wilson, James. (2008) *Moravians in Prague: A sociolinguistic study of dialect accommodation in the Czech Republic*. Ph.D. dissertation, University of Sheffield.
- Yilmaz, Gülsen and Monika S. Schmid. (2019) "First language attrition and contact linguistics". Jeroen Darquennes, Joe Salmons, and Wim Vandebussche, eds. *Language contact: An international handbook*. Vol. 1. Berlin: De Gruyter Mouton, 198–209. [Handbooks of Linguistics and Communication Science, vol. 45, no.1]
- Zajícová, Lenka. (2009) "Grammatical changes in Czech spoken by the immigrant community in Paraguay". Bohumil Vykypěl and Vít Boček, eds. *Recherches fonctionnelles et structurales*. Munich: Lincom Europa, 139–50.
- . (2012) "Language contact, language decay, and morphological change: Evidence from the speech of Czech immigrants in Paraguay". Martine Vanhove, Thomas Stolz, Aine Urdze, and Hitomi Otsuka, eds. *Morphologies in contact*. Berlin: Akademie Verlag, 283–307.
- Zíková, Magdalena. (2017) *Gramatikalizační potenciál anaforické funkce lexému ten v mluvených narativách* [Grammaticalization potential of the anaphoric *ten* in spoken narrative discourse]. Ph.D. dissertation, Charles University.

Chloe Castle
 Department of Language and Culture
 UiT The Arctic University of Norway
 Tromsø, Norway
 chloe.castle@uit.no

Appendices

Appendix 1: ABS Statistics – Czech Ancestry in SA (Choice 1)

For the group who chose Czech ancestry as choice 1, the following parameters were selected:

- 2016 Census > Cultural Diversity

COLUMNS: > Geographical areas (Usual Residence) > Main Statistical Area Structure > South Australia (State UR)

ROWS: > Selected Person Characteristics > ANC1P Ancestry 1st response > Southern and Eastern European > Eastern European > Czech

The results generated by TableBuilder are reproduced below.

Australian Bureau of Statistics		
2016 Census - Cultural Diversity		
ANC1P - 4 Digit Level by STATE (UR)		
Counting: Persons Place of Usual Residence		
Filters:		
Default Summation	Persons Place of Usual Residence	
STATE (UR)	South Australia	Total
ANC1P - 4 Digit Level		
Czech	870	870
Total	870	870
Data source: Census of Population and Housing, 2016, TableBuilder		
INFO Cells in this table have been randomly adjusted to avoid the release of confidential data. No reliance should be placed on small cells.		
Copyright Commonwealth of Australia, 2021, see abs.gov.au/copyright		
ABS data licensed under Creative Commons, see abs.gov.au/ccby		

Appendix 2: ABS Statistics – Czech Ancestry in SA (Choice 2)

For the group who chose Czech ancestry as choice 2, the following parameters were selected :

- 2016 Census > Cultural Diversity

COLUMNS: > Geographical Areas (Usual Residence) > Main Statistical Area Structure > South Australia (State UR)

ROWS: > Selected Person Characteristics > ANC2P Ancestry 2nd response > Southern and Eastern European > Eastern European > Czech

The results generated by TableBuilder are reproduced below.

Australian Bureau of Statistics		
2016 Census - Cultural Diversity		
ANC2P - 4 Digit Level by STATE (UR)		
Counting: Persons Place of Usual Residence		
Filters:		
Default Summation	Persons Place of Usual Residence	
STATE (UR)	South Australia	Total
ANC2P - 4 Digit Level		
Czech	809	809
Total	809	809
Data source: Census of Population and Housing, 2016, TableBuilder		
INFO Cells in this table have been randomly adjusted to avoid the release of confidential data. No reliance should be placed on small cells.		
Copyright Commonwealth of Australia, 2021, see abs.gov.au/copyright		
ABS data licensed under Creative Commons, see abs.gov.au/ccby		

Appendix 3: ABS Statistics for Figure 1 – Czech Ancestry in SA (Choice 1)

For the group who chose Czech ancestry as choice 1, the following parameters were selected:

- 2016 Census > Cultural Diversity

COLUMNS: > Local Government Areas (2016 Boundaries) (UR) > South Australia (LGA (UR))

ROWS: > Selected Person Characteristics > ANC1P Ancestry 1st response > Southern and Eastern European > Eastern European > Czech

The results generated by TableBuilder are reproduced below. The map function was then used to create the maps in Figure 1. This function is no longer available in TableBuilder.

Australian Bureau of Statistics

2016 Census - Cultural Diversity

LGA (UR) by ANC1P - 4 Digit Level

Counting: Persons Place of Usual Residence

Filters:

Default Summation

Persons Place of Usual Residence

ANC1P - 4 Digit Level LGA (UR)	Czech	Total
Adelaide (C)	8	8
Adelaide Hills (DC)	25	25
Alexandrina (DC)	10	10
Anangu Pitjantjatjara (AC)	0	0
Barossa (DC)	4	4
Barunga West (DC)	0	0
Berri and Barmera (DC)	0	0
Burnside (C)	25	25
Campbelltown (C)	20	20
Ceduna (DC)	0	0
Charles Sturt (C)	60	60

Clare and Gilbert Valleys (DC)	0	0
Cleve (DC)	0	0
Cooper Pedy (DC)	3	3
Copper Coast (DC)	0	0
Elliston (DC)	0	0
Flinders Ranges (DC)	0	0
Franklin Harbour (DC)	0	0
Gawler (T)	5	5
Goyder (DC)	4	4
Grant (DC)	0	0
Holdfast Bay (C)	41	41
Kangaroo Island (DC)	0	0
Karoonda East Murray (DC)	0	0
Kimba (DC)	0	0
Kingston (DC)	0	0
Light (RegC)	4	4
Lower Eyre Peninsula (DC)	3	3
Loxton Waikerie (DC)	0	0
Mallala (DC)	3	3
Maralinga Tjarutja (AC)	0	0
Marion (C)	67	67
Mid Murray (DC)	0	0
Mitcham (C)	49	49
Mount Barker (DC)	10	10
Mount Gambier (C)	3	3
Mount Remarkable (DC)	0	0
Murray Bridge (RC)	4	4
Naracoorte and Lucindale (DC)	4	4
Northern Areas (DC)	0	0
Norwood Payneham St Peters (C)	23	23
Onkaparinga (C)	129	129
Orroroo/Carrieton (DC)	0	0
Peterborough (DC)	0	0
Playford (C)	25	25

Port Adelaide Enfield (C)	61	61
Port Augusta (C)	3	3
Port Lincoln (C)	4	4
Port Pirie City and Dists (M)	0	0
Prospect (C)	6	6
Renmark Paringa (DC)	4	4
Robe (DC)	0	0
Roxby Downs (M)	5	5
Salisbury (C)	98	98
Southern Mallee (DC)	0	0
Streaky Bay (DC)	0	0
Tatiara (DC)	4	4
Tea Tree Gully (C)	60	60
The Coorong (DC)	0	0
Tumby Bay (DC)	0	0
Unley (C)	26	26
Victor Harbor (C)	5	5
Wakefield (DC)	4	4
Walkerville (M)	4	4
Wattle Range (DC)	0	0
West Torrens (C)	28	28
Whyalla (C)	5	5
Wudinna (DC)	0	0
Yankalilla (DC)	3	3
Yorke Peninsula (DC)	4	4
Unincorporated SA	4	4
No usual address (SA)	0	0
Migratory - Offshore - Shipping (SA)	0	0
Total	870	870

Data source: Census of Population and Housing, 2016, TableBuilder

INFO Cells in this table have been randomly adjusted to avoid the release of confidential data. No reliance should be placed on small cells.

Copyright Commonwealth of Australia, 2021, see abs.gov.au/copyright
ABS data licensed under Creative Commons, see abs.gov.au/ccby

**Appendix 4: ABS Statistics for Figure 1 – Czech Ancestry in SA
(Choice 2)**

For the group who chose Czech ancestry as choice 2, the following parameters were selected:

- 2016 Census > Cultural Diversity

COLUMNS: Local Government Areas (2016 Boundaries) (UR) > South Australia (LGA (UR))

ROWS: > Selected Person Characteristics > ANC2P Ancestry 2nd response > Southern and Eastern European > Eastern European > Czech

The results generated by TableBuilder are reproduced below. The map function was then used to create the maps in Figure 1. This function is no longer available in TableBuilder.

Australian Bureau of Statistics		
2016 Census - Cultural Diversity		
LGA (UR) by ANC2P - 4 Digit Level		
Counting: Persons Place of Usual Residence		
Filters:		
Default Summation	Persons Place of Usual Residence	
ANC2P - 4 Digit Level	Czech	Total
LGA (UR)		
Adelaide (C)	19	19
Adelaide Hills (DC)	28	28
Alexandrina (DC)	13	13
Anangu Pitjantjatjara (AC)	0	0
Barossa (DC)	3	3
Barunga West (DC)	5	5
Berri and Barmera (DC)	0	0
Burnside (C)	21	21
Campbelltown (C)	28	28
Ceduna (DC)	0	0
Charles Sturt (C)	44	44

Clare and Gilbert Valleys (DC)	3	3
Cleve (DC)	0	0
Cooper Pedy (DC)	11	11
Copper Coast (DC)	3	3
Elliston (DC)	0	0
Flinders Ranges (DC)	0	0
Franklin Harbour (DC)	0	0
Gawler (T)	14	14
Goyder (DC)	0	0
Grant (DC)	5	5
Holdfast Bay (C)	20	20
Kangaroo Island (DC)	0	0
Karoonda East Murray (DC)	0	0
Kimba (DC)	0	0
Kingston (DC)	0	0
Light (RegC)	13	13
Lower Eyre Peninsula (DC)	5	5
Loxton Waikerie (DC)	0	0
Mallala (DC)	5	5
Maralinga Tjarutja (AC)	0	0
Marion (C)	50	50
Mid Murray (DC)	0	0
Mitcham (C)	41	41
Mount Barker (DC)	26	26
Mount Gambier (C)	13	13
Mount Remarkable (DC)	0	0
Murray Bridge (RC)	6	6
Naracoorte and Lucindale (DC)	3	3
Northern Areas (DC)	0	0
Norwood Payneham St Peters (C)	27	27
Onkaparinga (C)	94	94
Orroroo/Carrieton (DC)	0	0
Peterborough (DC)	0	0
Playford (C)	28	28

Port Adelaide Enfield (C)	42	42
Port Augusta (C)	0	0
Port Lincoln (C)	7	7
Port Pirie City and Dists (M)	0	0
Prospect (C)	9	9
Renmark Paringa (DC)	4	4
Robe (DC)	0	0
Roxby Downs (M)	0	0
Salisbury (C)	49	49
Southern Mallee (DC)	0	0
Streaky Bay (DC)	0	0
Tatiara (DC)	3	3
Tea Tree Gully (C)	48	48
The Coorong (DC)	4	4
Tumby Bay (DC)	0	0
Unley (C)	27	27
Victor Harbor (C)	5	5
Wakefield (DC)	0	0
Walkerville (M)	4	4
Wattle Range (DC)	0	0
West Torrens (C)	38	38
Whyalla (C)	14	14
Wudinna (DC)	0	0
Yankalilla (DC)	0	0
Yorke Peninsula (DC)	0	0
Unincorporated SA	0	0
No usual address (SA)	0	0
Migratory - Offshore - Shipping (SA)	0	0
Total	809	809

Data source: Census of Population and Housing, 2016, TableBuilder

INFO Cells in this table have been randomly adjusted to avoid the release of confidential data. No reliance should be placed on small cells.

Copyright Commonwealth of Australia, 2021, see abs.gov.au/copyright
 ABS data licensed under Creative Commons, see abs.gov.au/ccby

Appendix 5: Diskusní Témata/Discussion Themes

Discussion themes were originally provided in Czech:

Cestování:

kde jste všude byli?
jaká místa chcete ještě navštívit?
v Austrálii, v ČR, jinde na světě

Život v České republice

Život v Austrálii

Filmy, které jste viděli v poslední době:

české filmy
americké filmy
australské filmy
filmy odjinud

Tři nejzajímavější věci, které jste kdy udělali

Oblíbená kniha nebo nejhorší kniha, kterou jste kdy četli

Oblíbené jídla nebo neoblíbené jídla, recepty, rozdíly mezi českou a australskou kuchyní

Co budete dělat o víkendu?

Jaké je vaše vysněné povolání/zaměstnání?

The translation is given below:

Travel:

where have you traveled to in the world?
what places do you want to visit?
in Australia, in the Czech Republic, elsewhere in the world

Life in the Czech Republic

Life in Australia

Films that you have seen recently:

Czech films
American films
Australian films
films from other countries

The three most interesting things you have ever done

Favorite book or worst book you have ever read

Favorite or least favorite food, recipes, differences between Czech and Australian cuisine

What are you doing on the weekend?
What is your dream job?

Appendix 6: Bilingual Ability Section of the Basic Information Form

Bilingual ability /dvojazyčné schopnosti:

English/Angličtina:											
0	1	2	3	4	5	6	7	8	9	10	
Czech/Čeština:											
0	1	2	3	4	5	6	7	8	9	10	

0 = does not speak the language at all/nemluví jazykem vůbec
10 = native-level fluency and maintained use of language/rodilý mluvčí a udržované používání jazyka

Appendix 7: Total Minutes Participant is Speaking in Observation Session

Participant	Minutes of participant speech within observation session	Total observation session time
Adéla	4 mins 5 seconds	15 mins 7 seconds
Dana	5 mins 16 seconds	9 mins 58 seconds
Eva	3 mins 48 seconds	9 mins 58 seconds
Jana	8 mins 37 seconds	14 mins 6 seconds
Ivana	3 minutes	14 mins 6 seconds
Kamila	6 mins 52 seconds	16 mins 21 seconds
Milada	3 mins 35 seconds	14 mins 6 seconds
Zuzana	8 mins 5 seconds	15 mins 7 seconds
Roman	3 mins 15 seconds	16 mins 21 seconds
Martin	2 mins 38 seconds	15 mins 7 seconds

Appendix 8: Presence of Participant Attestations in the Czech National Corpus (SYN2020) (Křen et al. 2020)

The following table presents the participant attestations discussed in this article, shows whether they are present in the SYN2020 subcorpus of the Czech National Corpus, and offers an explanation or comparison of attestations' appearance or non-appearance in the CNC. This gives some insight as to whether the attestations produced by participants occur in Czech as spoken in the Czech Republic. The focus of the attestation is highlighted in bold (as it is in the main text). The KonText interface was used to search the corpus, to allow for the specific phrases to be searching using the "word" attribute.

Sections are shaded grey if they represent an example (to aid in the description of a certain grammatical phenomenon) rather than data.

Attestation	In Czech National Corpus (SYN2020)?	Explanation/Comparison
1		
2	<i>my jsme si to projeli,</i> my se podiváme	Used more frequently without pronoun Unconventionality more due to repetition of pronoun in subsequent phrase
3	já musím jet domů	Used more frequently without pronoun
4	takže místo tu rodinu v tam	Used much more frequently without preposition
5	osobní třídy čtvrtek-Ø	<i>třídý</i> occurs rarely in corpus (5 hits, 0.04ipm), alternate possibility to test <i>čtvrtek</i> without a preposition = <i>jít</i> (occurs 17,402 times in corpus, 142.84ipm), still no results found for <i>jít čtvrtek</i>
6	na sobotu z neděli	No results for this phrasing, but results for conventional phrasing (though few)
7	mám manžel-Ø a dcera	No results for unconventional (nominative) phrasing, but results for (accusative) conventional form
8	jíme neměli rodinu v Austráliie	No results for unconventional phrasing (nominative), but many more results for conventional form (locative)

* Instances per million words.

9	kvůli rodinu	kvůli rodinu = no results kvůli rodině = 33 hits, 0.27ipm	No results for unconventional phrasing (accusative), but results for conventional form (dative)
10	s pan-Ø učitelem	s pan učitelem = no results s panem učitelem = 17 results, 0.14ipm	No results for unconventional phrasing (nominative), but results for conventional form (instrumental)
11			
12	vím že moje brácha	moje brácha = no results můj brácha = 66 results, 0.54ipm	No results for unconventional phrasing (feminine gender), but results for conventional form (masculine gender)
13	to jsou moje lidi	moje lidi occurs 10 times	Each of the times <i>moje lidi</i> occurs, it is in the accusative case (whereas Jana uses it in a nominative position)
14	mám manžela a dvě kluky	dvě kluky = no results dva kluky = 52 hits, 0.43ipm	No results for unconventional phrasing (feminine/neuter gender), but results for conventional form (masculine gender)
15			
16	jsem si učila	jsem si učila = no results jsem se učila = 52 hits, 0.43ipm	No results for unconventional phrasing (<i>si</i>), but results for conventional form (<i>se</i>)
17	snažím Ø ted'ka	snažím ted'ka = no results snažím se ted'ka = also no results	No results for <i>snažím se ted'ka</i> , but context may be too specific. 8 hits for <i>snažím</i> , and each of them include <i>se</i> (0.07ipm)
18	jak se může říct	jak se může říct = no results	Issue here is contextual rather than grammatical correctness/frequency of use

19	on nechce jít camping	<i>jít camping</i> = no results <i>jít kempovat</i> = 1 result, 0.01 of corpus <i>kempovat</i> = 24 results, 0.2ipm <i>kempink, camping, kemping</i> = 32 hits, 0.26ipm no <i>jít kempink, camping, kemping</i>	Much rarer to use more analytic <i>jít camping, kempink, camping, kemping, kempovat</i> than to use <i>kempovat</i> or <i>kempink, camping, kemping</i> as a standalone
20	tam můžeš dělat i camping	<i>dělat i camping</i> = no results <i>dělat kempovat</i> = no results no <i>dělat kempink, camping, kemping</i>	No results for this construction
21	kdy vase děti šli do školky tady?	<i>kdy vase děti šli do školky tady</i> = no results	Phrase too long to test in corpus (unlikely this exact sentence or one comparable has been said/written). No results, in any case.
22	to bylo takové ty koule, to jsou ty české	<i>ty koule</i> = 13 hits, 0.11ipm <i>koule</i> = 2,051 hits, 16.84ipm <i>ty české</i> = 52 hits, 0.43ipm <i>české</i> = 34, 766 hits, 285.37ipm	Occur much more frequently without the demonstrative/tentative article
23	mám ty vnoučata	<i>ty vnoučata</i> = no results <i>vnoučata</i> = 518 hits, 4.25ipm	Occur much more frequently without the demonstrative/tentative article

Zone-Flooding as a Discursive Strategy of Czech Anti-System News Portals

Masako U. Fidler and Václav Cvrček

Abstract: The anti-system media (ANTS)—known for spreading disinformation—might seem to “flood [the media] zone” with a chaotic multitude of information: truths, untruths, and half-truths alike. The main goal of this study is to find evidence of systematicity in this seeming chaos: persistent and recurring narrative lines that run through the media class irrespective of the news topic. Two empirical methods (Keyword Analysis and Market Basket Analysis) are applied to large data from Czech online media (all articles, regardless of topic, from 40 ANTS web portals over three months in 2020). ANTS’ narratives are advanced by creating specific associations. The current approach is based on the idea that texts can be characterized with the help of conceptual associations, pursuing concepts which co-occur within the same text regardless of sentence or paragraph boundaries. This approach thus differs from the frequently-used strategy in discourse analysis of examining phenomena such as collocations, use of passive voice, or nominalization. The distinct properties of ANTS can be highlighted by contrasting it to the mainstream media class and to reader expectations in journalistic practice. The results, culled from servers including those not explicitly sponsored by the Kremlin, indicate that a schematic set of narrative lines permeate ANTS: a model of the world divided into the West (USA, NATO, and the EU) and Russia, in which the West has a negative image relative to that of Russia. These narrative lines lead to an argumentation for Czechia’s separation from the West (Czexit, leaving NATO) and for alignment with Russia.

1. Introduction*

This paper explores the discursive strategy of “flooding the [media] zone” in Czech anti-system media. According to the Oxford English Dictionary, the primary meaning of the transitive verb *to flood* is “to inundate”, and “to cover or fill with water; to irrigate [...]; to deluge with water”. Zone-flooding thus

* The following abbreviations are used throughout this article: AA (associative array), ANTS (Anti-system media), ANTS-AA (Anti-system media associative array), KW (keyword), KWA (keyword analysis), MBA (Market Basket Analysis), MS (Mainstream media), MS-AA (mainstream media associative array), and RefC (reference corpus).

refers in its literal sense to the act of filling (and indeed overfilling) a zone with water. The word is used also figuratively in a cybersecurity context: using massive amounts of traffic to block access to servers or websites (York 2010). In this paper, zone-flooding is understood as a discourse practice or strategy. It refers to the dissemination of information within the media zone and subsequently the public discourse, whether true or not, in massive quantities, especially in internet articles by anti-system media; its aim is said to disorient the public (Illing 2020).

The study aims to provide an empirical account of zone-flooding based on a representative (“topic-blind”) sample of data (§2): more specifically, it attempts to demonstrate the **consistency** with which anti-system media ties together apparently disparate topics to weave a recurrent set of underlying narratives via discourse framing, an extension of conceptual framing introduced by Fillmore (1982) on the level of lexical semantics.

The approach used in this paper is distinct from what is often pursued in Critical Discourse Analysis, such as stylistic and semantic aspects of usage and obfuscation of agency—e.g., use of nominalization and passive voice (Fairclough 2003: 12–13, 145–50). It is also different from grammar-based quantitative approaches that probe the implicit creation of image of social actors (e.g., Keymorph analysis by Fidler and Cvrček 2018, Cvrček and Fidler 2019, and Janda et al. 2022). The goal of this paper is to identify a stable set of overarching narratives created by means of conceptual associations which are repeated over and over in a large number of texts (cf. §2).¹ Such associations are likely to have incremental cognitive effects on readers’ interpretation of not only current but also newly arising situations. As we try to capture recurrence of associations, neither an analysis of a sizable corpus on a single topic (e.g., migration, racism, or COVID-19), nor an analysis of a small number of texts would prove the presence of such a phenomenon. It was necessary to apply methods that were shown to facilitate conceptual framing extraction (Cvrček and Fidler 2022) from a large volume of data. Presentation of text samples therefore is used not as a proof, but as an illustration of what is suggested by the quantitative results.

Sections 2 and 3 introduce the data and methodology. The results are discussed in §4, followed by conclusions in §5. Due to the large number of results, the pertinent data and the code used are available in an OSF repository (<https://osf.io/mkbzg/>).

¹ These narratives are akin to what George Lakoff refers to as “strict father model” (2004), which is an underlying conceptual schema instantiated by various arguments and policies supported by the US Republican Party.

2. Data

The target of our research is a corpus of texts (articles, including news, op-eds, interviews, and other journalistic forms) published by media portals that present themselves as providing “alternative views” not covered (or deliberately omitted) by the mainstream media, and which can be clustered on the basis of the similarity of their audience.

2.1. ANTS and MS Media Classes

The data were drawn from the ONLINE1 corpus (Cvrček and Procházka 2020), a monitoring corpus of online Czech, which has been created within the Czech National Corpus project. Our focus is on the data which cover the period from 1 June to 1 September 2020. This time span was intentionally chosen (at the time the research was conducted in 2021) for its topic diversity: it is situated between the end of the first wave of COVID pandemics and the beginning of the second wave in Czechia, where coronavirus was present but was not the “only” topic.

As the focus of this study is the discursive practice of an entire media class (approximately 40 web portals), the target corpus includes texts on varying topics, e.g., the migration crisis in Europe, the Black Lives Matter (BLM) movement, the Belarus protests, the Beirut port explosion, domestic news within Czechia (politics, train crashes, etc.) and news about the EU. This approach clearly differs from theme-based discourse studies, such as the focus of Baker and McEnery on the representation of refugees and asylum seekers in the UK press (Baker and McEnery 2005), of Islam and Muslims (Baker et al. 2013), of gay men (Baker 2005) or of genre-based discourse, such as the analysis of Czech presidential New Year’s addresses by Fidler and Cvrček (2019).

Not all texts were included in this study, but the only criteria for exclusion were formal: their size and the number of prominent units (keywords) within a text (cf. §3.1), since text length may affect the way in which we identify associations between prominent units (for details see §3). Texts of extraordinary length for an online newspaper article (short texts with less than 500 tokens and large texts with more than 5,000 tokens) were thus excluded from the dataset. Additionally, only texts with at least 15 different prominent units/topics were retained, as a smaller number of keywords indicates a thematically opaque text without enough surface area for the study of topic interrelation.

Our media type classification follows the ONLINE corpus annotation, which is based on Josef Šlerka’s audience-based typology of Czech news portals. Šlerka (2018) uses online reader behavior patterns, such as visits to websites (based on Alexa Rank; <https://www.alexa.com/>) and sharing and liking social media articles (based on the CrowdTangle service). Such information

allows for the creation of media site clusters irrespective of their linguistic characteristics, topic preferences, or political stance, based solely on a significant overlap in audience. The only part of the classification that is subject to researcher interpretation is the labeling of each cluster, which is derived from the features of a particular web portal that can be considered a cluster prototype.

The current study uses three of Šlerka's media clusters. Our target media class comprises the "Anti-system" cluster and a related cluster labeled "Political tabloid"; we will refer to this group as ANTS as both typically challenge the mainstream media and spread disinformation.² The third media class, used as the control dataset, is drawn from Šlerka's "Mainstream" media cluster; because of its disproportionally large size relative to the ANTS portals, the number of "Mainstream" portals was reduced by keeping only the 11 most important ones (in terms of the number of visitors and the number of texts)³; this group of portals is referred to as MS. The resulting corpus (i.e., both MS and ANTS) has the following parameters:

Table 1. Target corpus

Media class	Number of texts	Number of words	Average number of words in a text	Standard deviation
ANTS	4,352	5,219,362	1199	674
MS	10,841	9,552,895	881	474

3. Methods

In this paper, we use two quantitative methods for analyzing the MS and ANTS subcorpora accompanied by qualitative inspection of text samples: we first use keyword analysis (KWA) to identify the prominent units/concepts

² The list of analyzed ANTS servers consists of political tabloids (gloobal.cz, ireporter.cz, irucz.ru, necenzurujeme.cz, parlamentnilisty.cz) and antisystem servers (ac24.cz, aeronet.cz, blogspot.com, casopis-sifra.cz, casopisargument.cz, ceskoaktualne.cz, czech.cri.cn, Czechfree-press.cz, duchdoby.cz, e-republika.cz, eportal.cz, ers.blog.cz, euportal.cz, eurabia.cz, euserver.cz, infokuryr.cz, isstras.eu, leva.net, levaperspektiva.cz, necenzurujeme.cz, novaburzoazie.com, novarepublika.cz, nwoo.org, orgo.net, pravyprostor.cz, protiproud.cz, rukojmi.cz, skrytapravda.cz, sputniknews.cz, stredoevropan.cz, svobodnenoviny.eu, vlasteneckenoviny.cz, zpravy.dt24.cz, zvedavec.org, and webnode.cz). Note that some of the URLs might not be accessible as a result of a decision taken by the Czech authorities in February 2022 to shut down some of the pro-Russian web portals after the start of Russian aggression on Ukraine. The texts are still available in the ONLINE corpus.

³ The list of major MS media consists of ceskenoviny.cz, denik.cz, idnes.cz, ihned.cz, impuls.cz, irozhlas.cz, lidovky.cz, nova.tn, novinky.cz, reflex.cz, respekt.cz.

in articles, and then investigate the co-occurrence of keywords within texts by Market Basket Analysis (MBA). Both methods are briefly described below.

3.1. Keyword Analysis of Texts

Keyword analysis (KWA), first coined by Mike Scott (Scott and Tribble 2006), is used by an increasing number of corpus-based discourse studies as a starting point of analysis (Baker 2006: 125). KWA compares the relative frequencies of words in the target text or corpus with the frequencies of the same words in the reference corpus (RefC), yielding keywords (KW), i.e., words that have statistically significantly higher frequency (measured in our case by log-likelihood test) in the target text than would be expected against the background of their frequency in the RefC. As several studies have found that statistical significance itself is not an appropriate way to measure keyness (Gabrielatos and Marchi 2012; Hofland and Johansson 1982), KWA is now often accompanied by effect-size estimators, in our case DIN (Fidler and Cvrček 2015: 204). KWs can reflect genre/register differences between text and the RefC and are said to point to the major topics of the target text or what the text is about (Scott 2010: 43). KWs in this study were identified with the following settings, separately for each text in ANTS and MS target subcorpora:

- Harvested prominent units: lemmas (only tokens with Czech letters)
- RefC: “offline” journalistic texts from 2015–2018 from SYNv8 corpus (Křen et al. 2019), tabloids and broadsheets, i.e., text of the same register as target texts.⁴
- Minimal frequency of a word in a text: 3 occurrences
- Test statistics: log-likelihood with 0.001 significance level
- Minimal KW DIN value in a text: 70 (max. value is 100)

⁴ The main goal of this study is to explore the underlying narratives, which are expected to impact readers’ cognition in the long term. Our interest in the connection to cognition motivates the use of RefC from 2015–18 journalistic texts. As this RefC is expected to reflect the stable journalistic linguistic usage that readers view as a norm, the harvested KWs using this RefC are expected to reflect concepts (lemmas) that are striking to the readers (cf. Fidler and Cvrček 2015 for discussion of KWA and RefCs) and not, e.g., the register differences. This approach yields both KWs indicating what is new in 2020 (news topics), and how they are interconnected. The alternative use of MS texts to yield KWs in ANTS and vice versa may seem appropriate, i.e., use of one subcorpus against the background of the other. This approach would illustrate how a specific group of readers view the target texts but would not highlight what might impact a wider, general group of readers. Furthermore, this approach would not allow a comparison of KWs between ANTS and MS as it would lack the same point of reference.

The thresholds were set based on our previous experience with similar analyses of Czech texts (Cvrček and Fidler 2019, 2022; Fidler and Cvrček 2018) and represent a conservative setting, only slightly prioritizing KW recall over precision.

3.2. Market Basket Analysis

As has been mentioned in previous studies, KWs are “just pointers” for interpretation (Scott 2010). An adequate and sound interpretation of KWs requires the context where these words typically occur. Moreover, since our goal is to characterize the ANTS media class as a whole, we must move beyond inspection of the role that KWs play in a single text and seek repeated KW co-occurrence in a large number of texts to reveal how KWs are systematically framed. The present study of KWs therefore applies Market Basket Analysis (MBA); for the efficacy of MBA and its detailed description, see Cvrček and Fidler 2022.

MBA is a data-mining technique used originally in marketing (Han et al. 2011; Information Resources Management Association 2014) to identify possible associations between items in shopping carts (one of the results may be suggestions such as “customers who bought X often buy Y”). MBA sifts through transactions in quantity, examines all possible combinations of items, and looks for associations among them, calculating three variables which describe the scope and strength of the link, which help filter out the unimportant ones. MBA applied to KWs in texts (considering texts as shopping carts and KWs as merchandise) utilizes the following three measurements:

- **Support:** probability of a text containing both KWs participating in the link (helps avoid associations occurring only in a limited portion of texts)
- **Confidence:** proportion of texts where KW *B* is present when it also contains KW *A* (filters out accidental associations)
- **Lift:** the strength of association—how much our confidence has increased that KW *B* will be present in an association given that KW *A* is already present

The threshold levels for MBA were set in the following manner:

- Confidence > 0.4
- As the number of texts in MS and ANTS differs, we adjusted the threshold level of support proportionately: ANTS support > 0.003, MS support > 0.0019
- Lift > 2.7

The threshold values were set empirically, based on previous studies, to keep the number of associative links manageable (for more details on the method and its technical parameters see Cvrček and Fidler 2022).

We obtained associative links among KWs up to the size of 4, i.e., with a maximum of three words preceding the arrow and one following it, e.g., *migrant*, *žadatel* ‘applicant’, *země* ‘country’ → *azyl* ‘asylum’, which can be read as “texts containing the words *migrant*, *applicant* and *country* as keywords will most probably also contain the word *asylum* as a keyword”. We identified 35,435 associative links in ANTS and 108,922 in MS using these settings.

To examine the entire pool of associations connected to one KW, we further refined our method to look at the associative array (AA) of a word, i.e., the entire set of associated KWs extracted from all associative links in which a particular KW is included. An AA of a KW contains associated KWs that help our understanding of how the KW is framed: e.g., the AA for the KW *organizace* ‘organization’ in MS connects the KW in summer 2020 to domestic politics, specifically a scandal with a local organization of the leading political party ANO⁵ (*Babiš*,⁶ *Brno*, *kauza* ‘affair’, *předsednictvo* ‘chairmanship’, *Vokřál*⁷), whereas the anti-system AA for the same KW includes associated KWs *terorista* ‘terrorist’, *válka* ‘war’, *protest* ‘protest’, *proti* ‘against’, *globální* ‘global’, *nadace* ‘foundation’, *Soros*,⁸ *Amerika* ‘America’, *BLM* (Black Lives Matter), connecting ‘organization’ to the BLM protests in the US, terrorism, and global NGOs.

As illustrated above, the pattern of framing (represented by AA) differs between the media classes, as was suggested by our previous small pilot study focusing on the single KW *migrant* ‘migrant’ (Cvrček and Fidler 2022). The present study covers multiple KWs from the entire target corpus with the aim of revealing ANTS’ distinct associations that point to its unique narrative lines. These narratives when repeated may in turn lead to a pervasive argumentation, regardless of news topics.

3.3. Contrast to Highlight Notable Features of the ANTS Media Class

The entire process to uncover how associative KW framing occurs, involving corpus compilation, KWA, and MBA, is summarized in Figure 1. The important

⁵ ANO ‘YES’, an acronym of the Czech populist party which stands for *Akce nespokojených občanů* ‘Action of dissatisfied citizens’.

⁶ The Czech prime minister and leader of the political party ANO.

⁷ A local politician within the political party ANO.

⁸ The billionaire hedge fund manager and philanthropist George Soros.

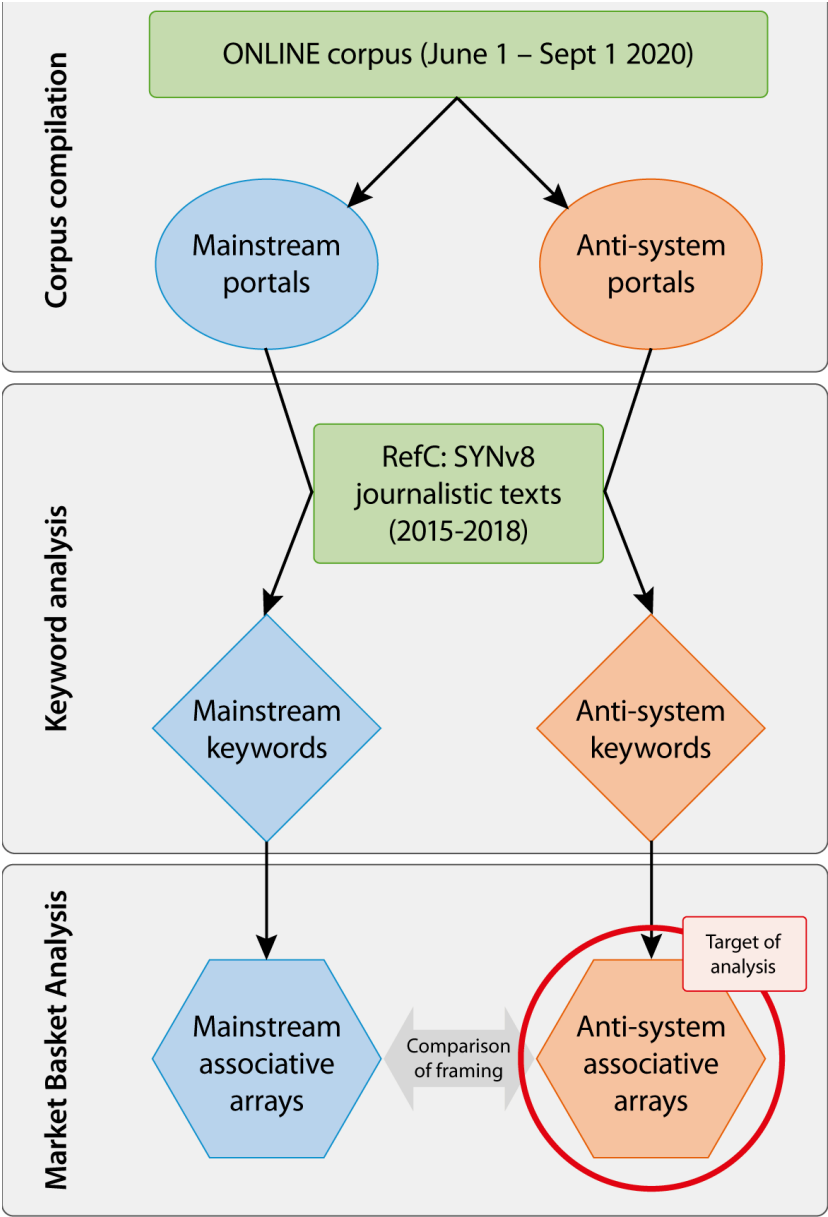


Figure 1. Progress of analysis

point here is that the unique features of the ANTS, the target of this study, are pursued in contrast to those of MS.⁹

In a nutshell, the approach in this study parallels the cognitive linguistic notions of frame and domain (Fillmore 1982) and “profile and base” (Langacker 1987: 147–82), which model our understanding of the meaning of a linguistic unit by using the notion of contrast.

3.4. Triangulation

Given the number of associations identified by MBA, too large to explore in its entirety, sampling was the only viable option. To increase the validity of our interpretations, we approached the data from three perspectives and triangulated the results. KWs and their associations were examined in three categories: media-class-dominant KWs, seasonal KWs, and shared KWs. They are not discrete groups and can overlap.

ANTS-dominant KWs (cf. §4.1) point to concepts preferred by one media class. A KW is “ANTS-dominant” when it meets two conditions:

- (a) it appears in at least 1% of ANTS texts as a KW;
- (b) the proportion of ANTS texts in which it appears as a KW is minimally twice as large as the proportion of MS texts in which it is also a KW.

For example, the word *prohlášení* ‘statement’ appears as a KW in 91 texts in ANTS and 62 texts in MS (which is 2.1% and 0.57%, respectively, cf. Table 1). The word is therefore dominant for ANTS since $0.021/0.0057 = 3.68$. We have identified 334 ANTS-dominant and 176 MS-dominant KWs.

Seasonal KWs (cf. §4.2) are expected to reflect topics that receive **short-term media attention**. They are defined here by three conditions:

- (a) a word appears as a KW in at least 10 texts within the target period (1 June–1 September 2020)
- (b) the relative proportion of texts where the word is a KW in the target period is minimally twice as large as the analogous proportion in the three-month period preceding and following the target period
- (c) the conditions (a, b) for the same KW are met in both ANTS and MS.

For example, the word *výbuch* ‘explosion’ appears as a KW in 27 ANTS texts and 83 MS texts during the target period; considering the number of texts in both

⁹ This is an operational decision to delineate the properties of ANTS without any evaluative judgment that MS is the “perfect” media class.

segments, the relative number of texts where the word appears is 19.6 texts per thousand in ANTS and 23.4 per thousand in MS. Given that relative numbers in adjacent periods are 0.78 (March–May) and 2.5 (September–December) for ANTS and 0.56 and 1.65 for MS, ‘explosion’ meets all three conditions (a, b, c) for being seasonal.

Seasonal KWs can be further divided into two groups with respect to the distribution of their associated KWs. Seasonal KWs in (i), below, are expected to reflect shorter-term issues that engage primarily one media class. Seasonal KWs in (ii) are the shorter-term issues that ANTS and MS frame differently.

(i) **Seasonal KWs predominantly framed by a single media class**
(cf. §4.2.1)

These are KWs for which only one media class has its own set of associated KWs. The other media class may have no associated KWs¹⁰ or may share some associated KWs with the first media class but does not have its own set of (specific) associated KWs.

(ii) **Seasonal KWs distinctly framed by each media class** (cf. §4.2.2)

These are KWs for which each media class has its own set of associated KWs. The media classes may or may not share some associated KWs. Either way, each media class is presumed to create a distinct set of media-specific associations for the same KW.

Shared KWs (cf. §4.3) are those KWs that appear in both ANTS and MS texts. They point to topics that were in the spotlight for both media classes and thus they are likely to reveal differences in framing.

The complete lists of KWs (divided into the categories described in this section) and their AAs are available in the OSF repository at <https://osf.io/mkbzg/>.

4. Results and Interpretation

The results show that all three groups of KWs and their AAs lead to a limited set of narrative lines regardless of news topic. Observations are made first on the basis of KWs and their associated KWs, followed by illustrative text samples.¹¹

¹⁰ The fact that a word is a KW does not entail its involvement in associations to other KWs.

¹¹ The citations may not contain all the associated KWs, since they may occur quite far (several sentences or even paragraphs) apart from one another within the same text.

4.1. ANTS-Dominant KWs

ANTS-dominant KWs directly suggest which concepts especially engage ANTS. They, unlike their MS counterparts, point to a polarizing discourse, which presents a simplistic or exaggerated view of situations and hints at alternative views. The linguistic features of polarization (Van Dijk 2016: 73–74) are present among ANTS-dominant KWs: the pronouns *my* ‘we’, *náš* ‘our’, and *jejich* ‘their’ are prominent, as are distancing devices such as the modifier *tzv.* ‘so-called’ and words related to identity, e.g., *národ* ‘nation’ or *národní* ‘national’. Words used in categorical statements and hyperbole (“intensification” in Reisigl and Wodak 2016: 33) further suggest a sweeping argumentation: the modifiers *všechen* ‘all’ and *všecek* ‘any kind of’, the emphatic adverbs or boosters *prostě* ‘simply’, *naprosto* ‘completely’, and the negative pronoun and adverbials *nic* ‘nothing’, *nikdy* ‘never’, and *nikoli* ‘by no means’. Indefinite quantification that could be used to overgeneralize is seen in the KWs *mnoho* ‘many’ and *mnohý* ‘many [of]’. The particle *přece* ‘after all’ marks what the ANTS considers to be obvious to the reader. The adverb *možná* ‘maybe’ and the particle *no* ‘well’ indicate doubt and hesitation to imply the existence of alternative “correct” views.

Tendencies to provide opinions and/or the ANTS’ version of “what really happened” can be observed in the ANTS-dominant KWs: nouns like *debata* ‘debate’, *diskuse* ‘discussion’, *názor* ‘opinion’, *myšlenka* ‘thought’, *odpověď* ‘answer’, *otázka* ‘question’, *pravda* ‘truth’, *víra* ‘belief’, *skutečnost* ‘reality’, and *lež* ‘lie’; the adverb *údajně* and particle *prý*, near-equivalents to ‘allegedly’; and the verb *zveřejnit* ‘to make public’. The polarizing linguistic devices mentioned above are not found among MS-dominant KWs.

The AA of the possessive pronoun for 1st person plural *náš* ‘our’ includes associated KWs *soběstačnost* ‘self-sufficiency’, *potravina* ‘food’, *EU*, and the populist party *SPD*¹², which by themselves suggest a protectionist and potentially anti-EU stance. Below is an illustrative example. The underlined words in bold style are the ANTS-dominant KWs and words in bold style are KWs associated with them. AAs are presented in parenthesis where the seed KW is separated by a colon and the associated KWs are linked by hyphens. As many of the original URLs are now blocked, we cite the text IDs in ONLINE1 corpus.

¹² *Svoboda a přímá demokracie* ‘Freedom and Direct Democracy’, a right-wing populist and nationalist party.

- (1) Title: *Czexit není cíl, nýbrž prostředek k dosažení cíle*

Excerpt: [...] *Právě nedávno EU v odpovědi na **naši** snahu o obnovu **naší** potravinové **soběstačnosti** naopak zvýšila podíl **potravin** k dovozu do ČR na 85%. (náš: soběstačnost-potravina-EU-SPD) (id: 2020-08-21_31_77)*

'Title: Czexit is not the goal but the means to reach the goal

Excerpt: [...] Just recently, in response to **our** efforts to restore **our** food **self-sufficiency**, the EU on the contrary increased the share of **food** to be imported into the Czech Republic to 85%.' (our: self-sufficiency-food-EU-SPD)

Prý 'allegedly' (particle) is associated with *evropský* 'European' and *unie* 'Union', which points to a doubtful or directly Eurosceptic stance, which can be illustrated in the following article and the text fragment:

- (2) Title: *Potravinově soběstačné Slovensko? Ani náhodou. Brusel zaslal vládě varování kvůli pokusu zvýšit prodej domácích výrobků*

Excerpt: *Eurokomise tvrdí, že zákon je v rozporu s právem **Evropské unie**. Podle komise jsou tím diskriminováni ostatní výrobci z EU. Omezuje to **prý** svobodu podnikání maloobchodníků s potravinami. Zahraniční řetězce varování ze strany **Evropské unie** vítají a žádají novou vládu, aby tento zákon změnila. Vedení EU dalo Slovensku 3 měsíce na to, aby reagovalo na toto varování. (prý: evropský-unie) (id: 2020-08-21_31_77)*

'Title: Food self-sufficient Slovakia? No way. Brussels has sent warning to government over attempt to increase sales of domestic products

Excerpt: The European Commission claims that the law is in conflict with **European Union** law. According to the Commission, it discriminates against other EU producers. It **allegedly** restricts the freedom of small food retailers to do business. Foreign chains are welcoming the warnings from the **European Union** and are asking the new government to change the law. The EU leadership has given Slovakia 3 months to respond to this warning.' (allegedly: European (adj)-union)

Note that the associated KWs do not necessarily occur in direct context in (1). In fact, *SPD* is not in this text fragment, but it occurs in a much larger context. In (2), the associated keywords can occur not only in the direct vicinity of the KW *prý*, but also in other parts of the text including the article title.

The following section will focus on the AAs of top ANTS-dominant KWs.

4.1.1. ANTS KW Framing

The top 10 ANTS-dominant KWs and their AAs can be contrasted to their MS counterparts to highlight ANTS-specific interests (tables 2 and 3).

Table 2 shows that MS tracks the increase of the COVID-infected and deaths, and the government measures to fight COVID (cf. KWs 3–6 with a large number of associated KWs including *koronavirus* ‘coronavirus’, *zemřít* ‘to die’, *nakažený/infikovaný* ‘infected’, and *počet* ‘number’ in their AAs) drawn also from major western sources (Reuters). The news was also culled from the period when the annual graduation and entrance school exams take place (‘this year’ is associated with ‘test-exam’). MS is seen to cover sports (cf. KWs 1, 7, 9, and 10, and their AAs). Such KWs associated with exams and sports are absent in ANTS-dominant KWs. In MS, COVID is associated predominantly with fact-tracking.

The top 10 ANTS-dominant KWs in Table 3 suggest a totally different emphasis from MS (KWs that did not yield associated KWs were omitted: *kapitalismus* ‘capitalism’, *onen* ‘that’).

It is possible to identify quite a different priority in ANTS against the background of the MS-dominant KWs. The ANTS-AAs suggest how the topics are framed. The KW ‘self-sufficiency’, associated with (*náš-potravina-zemědělství* ‘our-food-agriculture’), e.g., confirm ANTS’s critical stance on the EU agricultural policy as a tool to break solidarity among the EU members, as was discussed in §4.1.

In framing the KW *BLM*, ANTS connects the anti-racist demonstrations in the US not only to the KW ‘black’, but also to the KW ‘white’, an associated KW not found in the MS-dominant AAs. ANTS also connects ‘elite’ to the US and Trump. The AAs for these two KWs both point to and emphasize the division in race and power in the US. The examples, (3–4), in fact both claim an impending crisis or revolution in a US that is said to be severely divided.

- (3) Title: *Rasová karta v amerických volbách, a ještě falešná? [...] Proč je barva rasy najednou v USA tak důležitá? [...] Na podzim dojde v USA k pokusu o neo-marxistickou listopadovou revoluci!*

Excerpt: *Bílá rasa je v procesu likvidace. Černoch zastřelil 5-letého bílého chlapce, v podezření je rasový **BLM** motiv a nenávisť k bělochům vyvolaná v televizi (BLM: bílý) (id: 2020-08-14_34_177240)*

‘Title: The racial card in the American elections, and moreover (a) false (one)? Why is the color of race suddenly so important in the USA? [...] In the fall, it will result in an attempt at a neo-Marxist November Revolution in the USA!

Excerpt: The **white** race is in the process of liquidation. A black man shot a five-year old **white** boy, a racist **BLM** motive and hatred towards whites triggered by TV are suspected’ (BLM: white (adj))

Table 2. Top 10 MS-dominant KWs and their AAs (complete list is available in the OSF repository).

MS-dominant KWs	Associated KWs in the AA (selection)
<i>sezona</i> 'season'	<i>liga-ligový-klub</i> 'league (noun)-league (adj)-club'
<i>letos</i> 'this year'	<i>test-zkouška</i> 'test-exam'
<i>potorzený</i> 'confirmed (adj)'	<i>agentura-celkem-člověk-denní-evidovat-infikovaný-koronavirus-milion-ministerstvo-mrtvý-nárůst-než-nově-o</i> <i>patření-oznámít-patient-pandemie-počet-případ-Reuters-zdravotnictví-země-zemřít</i> 'agency-in total-person-daily-epidemic-to record-infected (adj)-coronavirus-million-ministry-dead (adj)-increase-than-newly-measure(s)-patient-pandemic-number-case-Reuters-healthcare-country-to die'
<i>Reuters</i>	<i>člověk-denní-infikovaný-informovat-koronavirus-milion-mrtvý-nákaza-pandemie-počet-potorzený-přibýt-</i> <i>případ-více-zaznamenat-země-zemřít</i> 'person-daily-infected (adj)-to inform-coronavirus-million-dead (adj)-infection-pandemic-num-ber-confirmed (adj)-to increase-case-more-to record-to die'
<i>Brazílie</i> 'Brazil'	<i>celkový-člověk-den-epidemie-koronavirus-milion-ministerstvo-mrtvý-nakažený-nárůst-než-počet-Reu-</i> <i>ters-více-zdravotnictví-zemřít</i> 'total (adj)-person-day-epidemic-coronavirus-million-ministry-dead-infected (adj)-in-crease-than-number-Reuters-more-healthcare-to die'

MS-dominant KWs	Associated KWs in the AA (selection)
<i>přibýt</i> 'to increase'	<i>agentura-celkem-člověk-evidovat-koronavirus-ministerstvo-nakažený-nárůst-než-ohnis-ko-opatření-počet-případ-rekordní-Reuters-rouška-test-více-dřavotnictví-zemřít</i> 'agency-in total-person-to record-coronavirus-ministry-infected-increase-than-epicenter-measure(s)-number-case-record (adj)-Reuters-mask-test-more-healthcare-to die'
<i>zápas</i> 'game'	<i>gól-minuta-off-střela</i> 'goal-minute-off-shot'
Ostrava	<i>ostravský</i> 'Ostrava (adj)'
<i>liga</i> 'league'	<i>klub-ligový-mistr-sezona-utkání</i> 'club-league (adj)-champion-season-match'
<i>minuta</i> 'minute'	<i>gól-míč-poločas-šance-střela-trenér-zápas</i> 'goal-ball-halftime-chance-shot-coach-game'

Table 3. Top 10 ANTS-dominant KWs and their AAs (complete list is available in the OSF repository).

ANTS-dominant KW	AAs
BLM	<i>Antifa-bílý-černoch-policie-rasismus-USA-všecek</i> 'Antifa-white (adj)-black (n)-police-racism-USA-all sorts of'
<i>Xaver</i> ¹	<i>ČT-moderátor-pořad-rada-Rada-radní-Šarapatka</i> ² <i>-televize-Veselý</i> 'ČT [Czech TV]-moderator-[TV] show-[ČT] council-council member-Veselý'
<i>Donbas</i>	<i>armáda-Bělorusko-informovat-koronavirus-Krym-luganský-Lukašenko-prohlásit-Putin-ruský-Trump-Ukrajina-USA-voják-Zelenský-zvěřit</i> 'army-Belarus-to inform-coronavirus-Crimea-luhansk-Lukashenka-to state-Putin-Russian (adj)-Trump-Ukraine-USA-soldier-Zelenskyy-to make public'
<i>Sputnik</i>	<i>Rusko-ruský</i> 'Russia-Russian'
SSSR 'USSR'	<i>Rusko-ruský-sovětský-USA-válka-vojenský</i> 'Russia-Russian-Soviet-USA-war-military'
<i>Kyjev</i> 'Kyiv'	<i>armáda-Donbas-informovat-koronavirus-Putin-ruský-ukrajinský-USA-voják</i> 'army-Donbas-to inform-coronavirus-Putin-Russian-Ukrainian (adj)-USA-soldier'

ANTS-dominant KW	AA's
Železný ³	ČT, moderátor-Svoboda ⁴ -televize 'ČT-moderator-Svoboda-television'
elita 'elite'	americký-Trump-USA 'American (adj)-Trump-USA'
ČT 'Czech TV'	český-člen-generální-médium-moderátor-názor-novinář-odtolání-politický-poplatek-pořad-poslanec- prezident-rada-radní-ředitel-Šarapatka-slovo-sněmovna-stížnost-Svoboda-veřejnoprávní-Veselý-vol- ba-zákon-Železný 'Czech-member-general (top level)-media-moderator-opinion-journalist-removal (laying off)-political-fee-program-parliamentarian-president-council-member-director-Šara- patka-word-parliament-complaint-Svoboda-law-public (TV) (adj)-Veselý-election-law-Železný'
soběstačnost 'self-sufficiency'	EU-komodita-náš-potravina-zemědělec-zemědělství 'EU-commodity-our-food-farmer-agriculture'

1. Luboš Xaver Veselý is a controversial moderator and journalist and a then-member of the ČT council. He openly supported the Trikolóra movement, which is known to be a staunchly Eurosceptic and national-conservative party.

2. Zdeněk Šarapatka was a member of the Czech TV council who fought against attempts by certain Czech political parties and Russian propaganda to influence the TV station.

3. Jakub Železný, a moderator and journalist. He is known to have criticized China openly in the Czech TV news commentaries.

4. Ludvík Svoboda was a Czech general who served in WWI and WWII on the Eastern front and was the socialist president during and after the Prague Spring and the Soviet invasion.

- (4) Title: [...] *Zvyšují pogromy Trumpovy šance na zvolení? Bezmezná nenávisť tzv. Demokratů. Příkopý dělicí společnost se mění v propasti. Kdo zradí? Občanská válka nejpозději v zimě?*

Excerpt: **Trump** se stal prezidentem v roce 2016 právě proto, že země již byla v hluboké krizi, do níž ji zavedly právě **elity** z obou parlamentních stran zapouzdřené do **amerického** establishmentu. (elita: americký-Trump-USA) (id: 2020-06-11_35_106667)

'Title: [...] Do pogroms increase **Trump's** chances of being elected? The so-called Democrats' boundless hatred. The trenches dividing the society are turning into abysses. Who will betray? A civil war no later than winter?

Excerpt: Trump became president in 2016 precisely because the country was already in a deep crisis, one that was led by **elites** from both parliamentary parties rooted in the **American** establishment.' (elite: American (adj)-Trump-USA)

The focus on a Cold-War-like competition between Russia and the US (not only involving Ukraine and Belarus, but also in relation to the pandemic) can be observed in the AAs for KWs *Donbas*, *SSSR* 'USSR', and *Kyjev* 'Kyiv', whose AAs contain both the US, Russia, Ukraine, Belarus, and coronavirus. The examples below illustrate a confrontation between the US and Russia into which Ukraine and Belarus are drawn.

- (5) *Operace tajných služeb USA a Ukrajiny, shrnul Putin události kolem zadržení Rusů v Bělorusku. (Donbas: Bělorusko-voják-vojenský-armáda-Ukrajina-USA) (id: 2020-08-29_28_170534)*

'US and Ukraine intelligence operations—Putin summarizes events surrounding the detention of Russians in **Belarus**.' (Donbas: Belarus-soldier-military (adj)-army-Ukraine-USA)

- (6) Title: *Američané začali na ukrajinských vojákách testovat vakcíny na Covid-19 jako na laboratorních krysách, ale dopadlo to špatně, v Charkově na tyto vakcíny již zemřelo 5 ukrajinských vojáků! (Kyjev: USA) (id: 2020-07-22_34_53080)*

'Americans began testing vaccines for Covid-19 on Ukrainian soldiers as well as on lab rats, but it went badly; in Kharkiv, 5 Ukrainian soldiers already died from these vaccines!' (Kyiv: USA)

ČT, the Czech mainstream public TV, as part of the West-oriented institution, is presented as being divided and chaotic. The AAs for KWs, [Luboš] *Xaver* [Veselý] and [Jakub] *Železný*, are moderators representing different worldviews. It is also noteworthy that the KW ČT is associated with 'complaint'.

These associated KWs indirectly question the reliability of the mainstream news station. Example (7) is consistent with what AAs suggest:

- (7) *Tady hrozí ČT obrovský průšvih. Pustili to živě, ale s chybou. A dnes to má řešit Rada ČT (ČT: stížnost) (id: 2020-07-24_248_146339)*
 'Here, ČT is in huge trouble. They broadcast it live, but with a mistake. And today, the Czech TV Council must deal with it.' (ČT [Czech TV]: complaint)

The negative image of the conniving US and the mainstream media can be contrasted with the implicitly positive image of Russia. *Sputnik*, referring to the vaccine, appears only in ANTS as a KW. Moreover, the persistent co-occurrence with KWs 'Russia' and 'Russian (adj)' suggests that the Sputnik V vaccine receives a special spotlight by virtue of being a **Russian** product (rather than from any other country)¹³; *Sputnik* is not associated with KWs referring to administering or testing the vaccine (e.g., government measures, patients). ANTS' positive stance on Sputnik V can be exemplified below:

- (8) Title: *Zájem o ruskou protikoronavirovou vakcínu se rozšiřují [sic] o další země Evropy a Asie (Sputnik: ruský) (id: 2020-08-14_41_177142)*
 'Interest in the **Russian** anti-coronavirus vaccine is spread [sic] to include other countries in Europe and Asia' (Sputnik: Russian (adj))

ANTS-dominant KWs and their AAs clearly point to a set of narratives: an anti-EU stance, a cold-war narrative where the US is presented negatively and Russia more positively, and a questionable role played by the mainstream media.

4.2. Seasonal KWs

Seasonal KWs are expected to reflect short-term topics. Those clustering in one media are expected to reflect that media's preference towards certain topics, while those occurring in both media are expected to show differing ways of framing the same KWs. Again, the aim of this comparison is to highlight how ANTS differs from MS. Seasonal KWs with AAs containing 15 or more associated KWs are discussed below (the threshold was set arbitrarily, due to space limitations, in order to discuss only the most connected KWs). The AAs of these seasonal KWs are not only consistent with the observations made

¹³ Note that this word is both prominent (overused) and measured for strength in association. The associated KW is therefore significant although it may seem natural for it to appear in the context of Sputnik.

in §4.1, but reveal in more detail how current events are treated by ANTS in contrast to MS.

4.2.1. Seasonal KWs Predominantly Framed by a Single Media Class

Seasonal KWs framed predominantly by MS point to domestic issues (e.g., accidents and crime): e.g., *nehoda* ‘accident’, *železnice* ‘railway’, and *vězení* ‘imprisonment’. KWs such as *Chorvatsko* ‘Croatia’, *Slovinsko* ‘Slovenia’, *červenec* ‘July’, and *srpen* ‘August’ suggest news on popular vacation destinations for Czechs. KWs such as *Agrofert*, *střet* ‘conflict [of interest]’, and *zájem* ‘interest’ suggest the abuse of EU subsidy by the conglomerate Agrofert, with ties to prime minister Babiš. The accident in Beirut was covered mainly by MS (*libanonský* ‘Lebanese’, *přístav* ‘port’, *výbuch* ‘explosion’). Below is an example with the KW ‘Agrofert’ and its AA (*střet-zájem-dotace-zákon*, ‘conflict-interest-subsidy-law’).

- (9) *Do ČR dorazily dvě předběžné zprávy EK, které se týkaly Babišova možného střetu zájmů. V první komise dospěla k závěru, že Babiš má dál vliv na Agrofert a současně jako premiér ovlivňuje použití peněz z EU. Druhá zpráva se týkala zemědělských dotací. České úřady následně zaslaly do Bruselu své reakce. (Agrofert: střet-zájem-dotace-zákon) (id: 2020-06-19_2_208198)*

‘Two preliminary reports from the EC [European Commission] concerning Babiš’s possible **conflict** of **interest** have reached the Czech Republic. The first commission concluded that Babiš continued to have influence on **Agrofert** and at the same time, as Prime Minister, influenced the use of EU money. The second report concerned agricultural **subsidies**. The Czech authorities subsequently sent their reactions to Brussels.’ (Agrofert: conflict-interest-subsidy-law).

Unlike its MS counterpart, seasonal KWs clustering in ANTS suggest topics that could polarize the public. They imply conflict (e.g., *rasistický* ‘racist’, *obvinění* ‘accusation’, *bomba* ‘bomb’, *násilí* ‘violence’, *nepokoj* ‘unrest’, *Black, Lives, Matter*) and social actors who stirred public discord (*Okamura*, *Kalousek*, *Jakeš*, and *Milada Horáková*).

The ANTS’ choice of the social actors may seem to be random, but on closer inspection, it is driven by overarching narratives against the West. Among the social actors, the AA for [Tomio] *Okamura* is most straightforward (*EU-SPD*). Okamura is known for his right-wing anti-immigration populist stance and is the leader of the populist SPD Party (Freedom and Direct Democracy). The associated KWs referring to the EU and SPD point to texts about Czechs exiting the EU, as seen, e.g., in (10):

- (10) *Pokud jde o Evropskou unii, SPD zastává názor, že je nereformovatelná, a pokud z ní nevystoupíme, tak nás zničí* (Okamura: EU-SPD) (id: 2020-07-06_31_131495)

‘As for the European Union, the **SPD** is of the opinion that it is unreformable, and if we do not leave it, it will destroy us’ (Okamura: EU-SPD)

[Miroslav] Kalousek’s actions can also be seen as an indirect link to the narrative against Czechia’s EU membership. Kalousek introduced strict austerity measures while acting as finance minister. He has often stirred controversy and public demonstrations have been organized against him. The associated KWs (*vláda* ‘government’-*Babiš-Andrej*) point to the ex-finance minister Kalousek’s critical stance on Andrej Babiš’s government and the growing state debt. Babiš is also associated with the misuse of EU subsidies (example 9). Coverage of Kalousek indirectly casts a negative light on the government, which works with the EU and so incurs economic difficulties.

- (11) *Exministr finance [Kalousek] následně konstatoval, z čeho pramení jeho přesvědčení, že Andrej Babiš je mafián. Skupina kolem premiéra prý kompletně “pohltila stát” a teď si sama nastavuje i vykládá pravidla hry a kdo to nevidí, je alibista.* (Kalousek: *vláda-Babiš-Andrej*) (id: 2020-06-17_22_204483)

‘The ex-finance minister [Kalousek] then stated the source of his belief that **Andrej Babiš** is a mafioso. The group around the prime minister is said to have completely “devoured the state” and now it sets and interprets the rules of the game all by itself; and whoever doesn’t see this is avoiding their responsibility.’ (Kalousek: *government-Babiš-Andrej*)

[Milada] Horáková and [Miloš] Jakeš are historical personalities from the communist period. Horáková is the only female western-oriented politician who was executed by the Communist government in the show trials in the 1950s and is a cultural icon of resistance against totalitarianism. The anniversary of her death stirred strong emotions and controversial debates. Jakeš is best known as the last General Secretary of the Communist Party of Czechoslovakia. His unprepared speech on 17 July 1989 is often viewed as exposing both his own intellectual incompetence and also the desperation of the communist regime.

Horáková is unsurprisingly associated with the KW *komunista* ‘communist’, reflecting the slogan ‘Executed by communists’, which was hung on many university buildings to commemorate the 70th anniversary of her death. In the text below, Horáková’s (alleged by ANTS) actions lead to an underlying

narrative: the unreasonable aggressiveness of Czech TV (represented by the TV moderator Jakub Železný) and the “fascination” over Horáková among the pro-EU and pro-NATO Prague liberals (pejoratively referred to as *Pražská havlérka* ‘Prague Havelites’).

- (12) *Jakub Železný začal na ČT vyhrožovat, že vyhodí ze studia kohokoliv, kdo se dotkne **Milady Horákové!** Česká televize už je zprivatizované médium, něco nám uniklo? A kdo doopravdy byla **Horáková**? Socialistka, poslankyně za ČSNS, hlasovala v jedné linii s **komunisty** Klementa Gottwalda, založila Svaz přátel SSSR, stala se místopředsedkyní svazu, v roce 1946 v Moskvě v hotelu National utajeně jednala s členy Nejvyššího politbyra ÚV KSSS a chtěla být ministryní zahraničí místo Jana Masaryka! Proč je Pražská havlérka tak fascinována ženou, která obdivovala národní socialismus a měla úzké vazby se Stalinovým politbyrem? (Horáková: komunista-Milada) (id: 2020-06-28_35_134693)*

‘Jakub Železný started threatening to kick out of the studio anyone who touches **Milada Horáková!** Is Czech TV already a privatized medium or did we miss something? And who was **Horáková** really? A socialist, an MP for the CSNS [Czech National Social Party]¹⁴, she voted in line with the **communists** of Klement Gottwald, founded the Union of Friends of the USSR, became vice-president of the union, in 1946 in Moscow at the Hotel National secretly negotiated with members of the Supreme Politburo of the Central Committee of the Soviet CP, and wanted to be foreign minister in place of Jan Masaryk! Why are the Prague Havelites so fascinated by a woman who admired National Socialism and had close ties with Stalin’s Politburo?’ (Horáková: communist-Milada)

Miloš Jakeš died in July 2020. As anticipated, Jakeš is linked to KSČ (The Czech Communist Party). The associated KW indicates not only Jakeš’s affiliation with communism, but his anti-West stance as well. The example below (a eulogy for him, as it were) elevates Jakeš to the status of a capable politician who “prophesied” that Czechia would be dictated to by the West (because of “Zionism sliding into Central and East Europe”).

- (13) *Jenže, v tomto známém projevu zazněly daleko důležitější informace a prorocké výroky, které tehdy v roce 1989 asi nikomu v hlavě nerezonovaly, ale při pohledu zpětně do doby před 31 lety je nyní hrozné zjištění, že generální tajemník ÚV KSČ už v červenci 1989 uměl popsat procesy nasunování*

¹⁴ The party should not be confused with the National Socialist Party led by Hitler. The article author apparently conflates this party with Horáková’s Czech National Social Party (social liberals).

sionismu do zemí Střední a Východní Evropy. **Miloš Jakeš** během projevu [sic] A tím hlavním procesem bylo zadlužování socialistických států takovým způsobem, aby jejich lidové vlády začaly dělat takovou politiku, jakou nadiktuje Západ. (Jakeš: Miloš-KSČ) (id: 2020-07-15_35_540)

‘But, in this well-known speech, there was much more important information and prophetic statements, which probably didn’t resonate in anyone’s head back in 1989, but looking back 31 years ago, we come to the terrible discovery that the general secretary of the Central Committee of **the KSČ** knew how to describe the processes of Zionism infiltrating the countries of Central and Eastern Europe as early as in July 1989. **Miloš Jakeš** during the speech [sic] And the main process was the socialist states falling into debt in such a way that their people’s governments begin to make the policies dictated by the West.’ (Jakeš: Miloš-KSČ)

Clearly, the four politicians above, even though they seem disparate, are used to further a goal of weaving the same line of narratives: narratives against the mainstream media, the Czech EU membership, i.e., essentially against the country’s western orientation. This underlying goal can be gleaned though by general knowledge of the individuals and their AAs, and has been illustrated by the text samples.

ANTS-AAs also suggest the racial polarization of the US society and problematizes protests against racism. While MS associates BLM-related KWs with very few, if any, concepts, the ANTS counterparts are consistently connected with not only *černý* ‘black’, but also *bílý* ‘white’. As seen in §4.1 with the KW *BLM*, the association array of *násilí* ‘violence’, e.g., leads to texts about the disadvantaged white population (14):

- (14) *Kdyby policista byl černý a pachatel bílý, řešil by to někdo? Záleží na bílém životě? (násilí: bílý-proti-černoch)* (id: 2020-08-02_38_101369)

‘If the policeman had been black and the offender **white**, would anyone care? Do **white** lives matter?’ (violence: white (adj)-against-black (n))

ANTS-KWs are connected to shorter-term events related to Russia, e.g., the AA for the KW (*díplomat: Rusko-BIS-vyhoštění* ‘diplomat: Russia-BIS [Czech Security Intelligence Service]-expulsion’). These associated KWs suggest news about the expulsion of Russian diplomats from Czechia, an action that is seen as unfair and victimizing Russia (cf. Fidler and Cvrček 2018: 217 and Cvrček and Fidler 2019: 105). The KW *ústava* ‘constitution’ is linked to the referendum on Russian constitutional changes; noteworthy is the somewhat unexpected inclusion of the KWs *USA* ‘USA’ and *americký* ‘American’ in the AA (*ústava:*

americký-Putin-Rusko-změna-ruský-USA ‘constitution: American (adj)-Putin-Russia-change-Russian (adj)-USA’). This AA indicates that the constitutional changes in Russia are important in relation to the US. The following example clarifies why the US is included: the old constitution is framed as “dictated by the US”, in line with the negative image observed in §4.1:

- (15) *Při nejmenším do doby, kdy se **ruští** občané rozhodnou nahradit v nouzi zavedenou ústavu diktovanou **USA** novou, vlastní. Ta bude, dojde-li k výměně ústavy, [...], definovat **Ruskou** federaci jako civilizační jednotku. **Rusko** jako civilizační jednotka je zcela něco jiného než století trvající chápání **Ruska** Západem. Taková **změna** nedovolí pokračovat v současném typu chování Západu vůči prezidentovi a **Ruské** federaci. (ústava: americký-Putin-Rusko-ruský (adj)-USA-změna) (id: 2020-06-30_61_48943)*

‘At least until such a time as **Russian** citizens decide to replace the constitution dictated by the **US** out of necessity with a new one, of their own. It will exist, if it results in replacement of the constitution, [...], [it will] define the **Russian** Federation as a unit of civilization. **Russia** as a unit of civilization is quite different from the West’s centuries-long understanding of **Russia**. Such a **change** will not allow the current type of behavior by the West toward the president and the **Russian** Federation to continue.’ (constitution: American (adj)-Putin-Russia-Russian (adj)-USA-change)

4.2.2. Seasonal KWs with Distinct Framing in Both Media Classes

Table 4 shows seasonal KWs in category (ii) with the number of associated KWs that are media-specific, and those in the intersection and the union of the two media classes. The KWs point mainly to the Belarus protests against presidential election results, the BLM movement in the US, and the EU summit. Since BLM has already been covered in the previous sections (§4.1 and §4.2.1), we will focus on the KWs and their AAs concerning Belarus and the EU here.

The MS-specific associations for the elections in Belarus, e.g., *Bělorusko: Cichanouská-agentura-výsledek* ‘Belarus: Tsikhanovskaya-[press] agency-result’, suggest that the mainstream media makes explicit references to the source of information (press agency) and focuses on the election results. It is also worth noting that one of the MS-specific associated KWs refers to the opposition leader and the presidential candidate against Lukashenko by name (‘Tsikhanovskaya’), unlike ANTS. The personal name humanizes the opposition and the protests, as can be also seen in the following example:

Table 4. Seasonal KWs for which each media class has its own set of associated KWs

KWs (gloss)	ANTS-specific associated KWs	MS-specific associated KWs	Intersection	Union
<i>Bělorusko</i> 'Belarus'	47	3	23	73
<i>protest</i>	36	4	32	72
<i>Lukašenko</i> 'Lukashenka'	45	4	20	69
<i>běloruský</i> 'Belarusian (adj)'	39	5	21	65
<i>policie</i> 'police'	22	5	9	36
<i>dotace</i> 'subsidy'	11	16	5	32
<i>policista</i> 'police officer'	10	5	12	27
<i>obnova</i> 'recovery'	11	4	11	26
<i>Minsk</i>	9	6	10	25
<i>policejní</i> 'police (adj)'	7	9	8	24
<i>demonstrant</i> 'demonstrator'	6	8	9	23
<i>Floyd</i>	9	4	10	23
<i>summit</i>	4	2	9	15
<i>Washington</i>	8	2	4	14
<i>rezoluce</i> 'resolution'	1	4	7	12

- (16) *Světů není dění v **Bělorusku** lhostejné. Některé země volají po nových volbách, jiné požadují sankce [...] Zároveň nabádala úřady, aby přistoupily na dialog, a vyzvala starosty, aby ve svých městech o víkendů zorganizovali poklidná masová shromáždění. **Cichanouská** také ohlásila záměr vytvořit koordinační radu pro předání moci. (Bělorusko: Cichanouská-agentura; id: 2020-08-14_7_8839)*

'The world is not indifferent to what is happening in **Belarus**. Some countries are calling for new elections, others are demanding sanctions [...] At the same time, she [Tikhanovskaya] has

urged the authorities to engage in dialogue and has called on mayors to organize peaceful mass rallies in their cities over the weekend. **Tikhanovskaya** has also announced her intention to create a coordination council for the transfer of power.’ (Belarus: Tikhanovskaya-agency)

In contrast to MS, ANTS-specific KWs are disproportionately numerous (cf. the numbers of ANTS-specific KWs for the Belarus-related KWs ‘Belarus’, ‘Lukashenka’, ‘Belarusian’, ‘Minsk’). These associations, however, are not randomly chosen. While both ANTS-AAs and MS-AAs include KWs referring to Russia (*Rusko* ‘Russia’, *Putin*), ANTS-specific KWs connect Belarus-related KWs to:

- the West, e.g., *západ* ‘West (n)’, *západní* ‘western’, *EU*, *NATO*, *USA*
- military- and security-related KWs, e.g., *bezpečnost* ‘security’, *voják* ‘soldier’, *vojenský* ‘military (adj)’, *armáda* ‘army’, *tajný* ‘secret [service]’.

ANTS is thus distinct from MS in pointing to the Cold War model—Russia vs. the West—to cover the current situation in Belarus. ANTS characterizes this relationship as “different shackles”, as shown in (17).

- (17) *Médiím v otázce **Běloruska** nelze věřit. Velice ohraný scénář. Co je VIP poukázkou na Majdan? Chtějí Bělorusové do jiných okovů? Kdo to asi všechno řídí? Zkušenosti máme habaděj* (Bělorusko: západ-západní) (id: 2020-08-14_248_42686)

‘The media cannot be trusted on the **Belarus** question. An out-of-date script. What’s a VIP voucher for the Maidan? Do the Belarusians want to get themselves into different shackles? Who’s likely to be masterminding it all? We have heaps of experience.’ (Belarus: west (n)-western)

ANTS also connects Belarus to NATO, Poland, and the EU, drawing special attention to the interaction between Belarus and the West, especially Poland as a NATO member sharing a state border with Belarus. This image of Belarus as a state that might be pulled towards the West and away from Russia by force has been observed also in §4.1.1 as well as in the example below. Russia is presented as a helper, reassuring Belarus it will send an army at the latter’s request.

- (18) *Andrej Babiš vyzval **EU** k podpoře státního převratu v **Bělorusku** podle modelu z Československa v roce 1989, běloruský prezident na to zareagoval hrozbou aktivace článku smlouvy o kolektivní bezpečnosti s Ruskou federací,*

Vladimir Putin potvrdil, že vyšle ruskou armádu do **Běloruska** po přijetí žádosti z Minsku! Na polské straně hranice se formují a hromadí tanky a vojska **NATO** v rámci cvičení Defender 2020, ale jako cvičení to nevypadá, běloruská vojenská rozvědka má obavy, že by mohlo dojít k pozvání vojsk do **Běloruska** ze strany běloruské opozice, kterou by předtím Brusel uznal za legitimní vládu v čele se samozvanou prezidentkou! (Bělorusko: Polsko-NATO-EU) (id: 2020-08-16_33_169750)

‘Andrej Babiš has called on the **EU** to support a revolution in **Belarus** modelled on Czechoslovakia in 1989, the Belarusian president responded with a threat to activate an article of a collective security agreement with the Russian Federation, and Vladimir Putin confirmed that he would send the Russian army to **Belarus** after accepting a request from Minsk! On the Polish side of the border, **NATO** tanks and troops are forming and massing as part of the Defender 2020 exercise, (but it doesn’t look like an exercise!), and Belarusian military intelligence is concerned that there may be an invitation of troops into **Belarus** by the Belarus opposition, which Brussels would acknowledge in advance as the legitimate government headed by a self-proclaimed president!’ (Belarus: Poland-NATO-EU)

The sense of concern about the West’s encroachment on Belarus can be observed via the AA that connects the Belarus protest with Ukraine, the locus of Euromaidan (*protest: Ukrajina* ‘protest: Ukraine’), an event that could trigger confrontation between the West and Russia.

- (19) *Běloruská tajná služba zachytila telefonické hovory z České republiky, které měly organizovat **protesty** v běloruských ulicích po zvolení Alexandra Lukašenka! Podle zdrojů běloruské tiskové agentury BelTA mělo jít o zaměstnance “státního média” a pracovníky české mezinárodní neziskové organizace bez uvedení konkrétních jmen nebo názvů! [...] V Bělorusku se schyluje k Majdanu, na ulicích v Minsku vyrůstají barikády, ale běloruský prezident varuje, že nepokoje v zemi organizují síly nejen ze Západu, ale i z Moskvy! (protest: Ukrajina-Rusko) (id: 2020-08-11_32_326)*

‘The Belarusian secret service has intercepted phone calls from the Czech Republic that they [telephone calls] are going to organize **protests** in the streets of Belarus after the election of Alexander Lukashenka! According to the Belarusian news agency BelTA, this was supposed to relate to employees of the [Czech] “state media” and workers of a Czech international NGO without mentioning specific [personal] names or [NGO] names! [...] Maidan is brewing in Belarus, barricades are being erected on the streets of Minsk, but the Belarusian president warns that unrest in the country is being

organized by forces not only from the West, but also from Moscow!’
(protest: Ukraine-Russia)

ANTS also diverges from MS with regards to the EU. The KWs *dotace* ‘subsidy’ and *obnova* ‘[EU COVID] recovery [plan]’ indicate that both media classes engage in covering EU economic policies as current topics. The ANTS-AA for the KW *obnova* ‘recovery’, however, includes associated KWs suggesting a burden on the member states (*dluh-krize-půjčka* ‘debt-crisis-loan’), unlike the MS-AA, which includes words expected in reports on EU negotiations (*dohoda-jednání-grant* ‘agreement-negotiation-grant’). The following ANTS example amplifies the state debt until it is comparable to the medieval plague:

- (20) *Občany, kteří by si na svých mobilech a tabletech něco četli na internetu, by Brusel zkasíroval novou digitální daní. Tohle není cesta z krize. Tohle je cesta do největší ekonomické krize od dob, kdy ve středověku třetina Evropanů zemřela na mor. (obnova: dluh-krize-půjčka) (id: 2020-06-01_43_57)*

‘Brussels would cash in on citizens browsing the Internet on their mobiles and tablets with a new digital tax. This is no way out of the **crisis**. This is the way into the biggest economic **crisis** since a third of Europeans died of the plague in the Middle Ages.’ (debt-crisis-loan)

4.2.3. Seasonal KWs: Summary

Seasonal KWs connected with current events show how ANTS advances its narrative of “Russia vs. the West” (the EU, NATO, USA, and pro-West mainstream media and NGOs). The KW framing in ANTS suggests repeated tendencies to look for multiple points of discord and dispute in the West, thereby placing an emphasis on the problematic aspects of the West.

4.3. Shared KWs (Non-Seasonal)

Shared KWs are perhaps the most informative in terms of ANTS’ distinct framing of the same KWs, in contrast to MS. The degree of overlap was measured by the Dice coefficient, which is calculated as the number of associated KWs in the overlap between media classes divided by their union. KWs with the lowest values (between 0 and 0.1), i.e., those which share no or a very small portion of associations, are listed below.

Dice = 0: *kauza* ‘affair’, *ředitel* ‘director’, *oběť* ‘victim’, *Praha* ‘Prague’, *banka* ‘bank’, *organizace* ‘organization’, *TOP* [09]¹⁵, *milion* ‘million’, *bezpečnostní* ‘of security’, *člen* ‘member’, *dluh* ‘debt’, *plán* ‘plan’, *finanční* ‘financial’, *pomoc* ‘help’

Dice < 0.1: *Turecko* ‘Turkey’, *turecký* ‘Turkish’, *Amerika* ‘America’, *kraj* ‘region’, *agentura* ‘agency’, *demokracie* ‘democracy’, *ministerstvo* ‘ministry’, *Řecko* ‘Greece’, *informovat* ‘to inform’, *nákaza* ‘infection’, *hranice* ‘border’, *vlak* ‘train’, *Německo* ‘Germany’, *případ* ‘case’, *armáda* ‘army’, *řecký* ‘Greek’ (adj), *počet* ‘number’, *nehoda* ‘accident’, *útok* ‘attack’, *unie* ‘union’, *německý* ‘German’ (adj), *test* ‘test’, *návrh* ‘proposal’, *doprava* ‘transportation’, *nakažený* ‘infected’, *USA*, *služba* ‘service’, *hnutí* ‘movement’, *karanténa* ‘quarantine’, *Rusko* ‘Russia’, *rouška* ‘mask’

Below we will focus on the country names that have large AAs: *Amerika* ‘America’ (*Amerika* was chosen instead of *USA* because the former’s lower Dice value), *Německo* ‘Germany’, *unie* ‘[European] Union’, *Turecko* ‘Turkey’, *Řecko* ‘Greece’, and *Rusko* ‘Russia’.

4.3.1. *Amerika* ‘America’

The ANTS-AAs for this KW are commensurate with the earlier observation about the US. ANTS links *USA* with violence and racial division (*Amerika: nepokoj-válka-černý-bílý-rasismus* ‘America: unrest-war-black-white-racism’), while MS links it with the North and South Americas and COVID (*Brazílie-latinský-koronavirus-nákaza-úmrť* ‘Brazil-Latin [America]-coronavirus-infection-death’). This link to pandemics can be found also in ANTS, but in addition ANTS-AA includes *flu*, a KW not found in the MS-AA. This associated KW suggests a representation of COVID-19 as “just a [strain of] flu”.

- (21) *Jak je to možné? Jednoduše: koronavirová **chřipka** prostě neexistuje. Odbyla si svoje, jako jedna ze tří chřipek, které se letos objevily na scéně a svůj vrchol měla v květnu. (Amerika: chřipka) (id: 2020-08-30_28_165134)*

‘How is that possible? Simple: Coronavirus **flu** just doesn’t exist. It’s done its bit, like one of three strains of flu that appeared on the scene this year and had its peak in May.’ (America: flu)

4.3.2. *Německo* ‘Germany’

The AAs of the shared KWs also reveal ANTS’ underlying narrative about the West. Unlike the MS-AA, the ANTS-AA includes *Soviet*, *Hitler*, and *war*. These

¹⁵ A Czech liberal-conservative political party (*Tradice Odpovědnost Prosperita*, ‘Tradition Responsibility Prosperity’).

associated KWs juxtapose and implicitly conflate Nazi Germany's past dominance with the current German dominance in the EU. In the example below, the Czech Republic is said to be "choking with anti-Russian rhetoric" and is turning into a "protectorate" of the "Germany-dominated EU":

- (22) *Tehdejší strach z **Německa** nás přivedl ke komunismu a ke čtyřiceti letům vazalství **Sovětskému** svazu, dnes se pro změnu zalykáme protiruskou rétorikou a nevíšimáme si toho, že se postupně znovu měníme v protektorát **Německem** dominované Evropské unie. (Německo: Hitler-sovětský-válka) (id: 2020-06-04_32_34927)*

'The fear of **Germany** at that time led us to communism and to forty years of vassalage to the **Soviet** Union; today we are choking with anti-Russian rhetoric for a change and are not aware that we are gradually turning again into a protectorate of the **German**-dominated European Union.' (Germany: Hitler-Soviet (adj)-war)

ANTS also criticizes "rewriting" the history of the liberation of Czechoslovakia from Nazi Germany:

- (23) *Oslavujeme Američany, vlasovce, nalháváme si, že jsme se osvobodili sami. Káci se pomník **sovětskému** vojevůdci a sundávají pamětní desky (Německo: Hitler-sovětský-válka) (id: 2020-06-04_32_34927)*

'We celebrate the Americans, the Vlasov¹⁶ army, deluding ourselves that we liberated ourselves. A monument to the **Soviet** military commander is being toppled' (Germany: Hitler-Soviet (adj)-war)

On the surface, (22) and (23) seem contradictory as the former presents the USSR differently: as a communist country that subjugated Czechoslovakia (22) and a country that liberated Czechoslovakia (23). What these texts have in common, however, is an indirect negative image of the West—the EU where Germany turns the Czech Republic into its vassal again and criticism of Czechia's pro-West inclinations and its disregard of the accomplishments of the USSR.

4.3.3. *Unie* '[European] Union'

The ANTS-AA of *unie* 'union', as in 'Germany', points to ANTS' critical narrative about the West. The associated KWs for 'union' in MS point to EU-Czechia interaction, policymaking, and negotiations (represented by prime minister

¹⁶ Andrei Vlasov, a Soviet Red Army general, collaborated with Nazi-Germany but helped Czechs during the Prague uprising at the end of WWII.

Babiš and his government): e.g., (*premier-Babiš-Česko-vláda-evropský* ‘prime minister-Babiš-Czechia-government-European’ (adj)). In contrast, ‘union’ in ANTS co-occurs with references to the Eurosceptic populist parties and to debt: (*trikolóra-SPD-dluh-půjčka* ‘Tricolor-SPD-debt-loan’) (see example 2). In addition, ‘union’ is associated with (*USA-NATO-Rusko-Ukrajina-Bělorusko* ‘USA-NATO-Russia-Ukraine-Belarus’), placing the EU among the geopolitical players involving Ukraine and Belarus. These KWs, as seen above, recur with many KWs in ANTS but are completely absent in the MS counterpart.

ANTS also differs from MS in terms of the number of associations with these KWs. Although these KWs are shared by both ANTS and MS, it is possible to state that the former “floods” the media space with much larger AAs. ANTS connects ‘union’ with 87 KWs and MS with 7. Similarly, ANTS connects *EU* with 155 KWs and MS with 20.

4.3.4. *Řecko* ‘Greece’ and *Turecko* ‘Turkey’

The ANTS-AAs for KWs ‘Greece’ and ‘Turkey’ suggest that they provide another opportunity for ANTS to bring up its persistent narrative regarding Russia. Associated KWs are again more numerous in ANTS than in MS: 50 KWs in ANTS and 9 in MS connected to ‘Greece’, and 59 KWs in ANTS and 1 KW in MS connected to ‘Turkey’. The MS-AA for ‘Greece’ points to summer vacation travel under COVID: (*cestovní-koronavirus-nákaza-srpen-test-turista* ‘travel-coronavirus-infection-August-test-tourist’). The ANTS-AA point to geopolitical conflicts: e.g., *NATO-EU-sankce-armáda-krize-Sýrie-Kypr-Rusko* ‘NATO-EU-sanctions-army-crisis-Syria-Cyprus-Russia’). Similarly, ANTS-AAs for ‘Turkey’ contain associated KWs referring to countries in conflict (*Sýrie-Kypr-Libye* ‘Syria-Cyprus-Libya’), social actors that are involved (*EU-NATO-Rusko*), contract and sanctions (*smlouva* ‘contract’, *sankce* ‘sanctions’), and military-related terms (*vojenský* ‘military’ (adj)). MS has only (*turecký* ‘Turkish’) in its AA. The following two examples are informative in terms of the role played by Russia. They contrast the EU’s incompetence and Russia’s important role in international conflict resolution.

- (24) Title: *Řecká vláda požádala Rusko o pomoc proti Turecku* (Řecko: *Turecko-Rusko-NATO-EU*) (id: 2020-07-23_32_179535)

‘The Greek government has asked **Russia** for help against **Turkey**’
(Greece: Turkey-Russia-NATO-EU)

- (25) *V březnu oznámilo Turecko úmysl již poněkolkáté vyslat do kyperských pobřežních vod těžební loď Yavuz, aby tam nelegálně těžila plyn. Kyperská vláda požádala o pomoc EU, která se však dodnes nedohodla na uvalení jakýchkoli účinných sankcí proti Turecku. Vláda neutrální kyperské*

republiky proto v nouzi požádala o pomoc Rusko (Turecko: Rusko-NATO-EU-Kypr) (id: 2020-08-23_27_89053)

‘In March, **Turkey** announced its intention to send the mining vessel Yavuz for the umpteenth time to Cypriot coastal waters for illegal gas drilling there. The Cypriot government has requested help from the EU, which, however, has not yet agreed to impose any effective sanctions against **Turkey** to this day. The government of the neutral Republic of Cyprus therefore out of necessity requested help from **Russia**.’ (Turkey: Russia-NATO-EU-Cyprus)

4.3.5. *Rusko* ‘Russia’

The ANTS-AA for ‘Russia’ shows the persistent narratives that place Russia center stage even more visibly. While MS connects *Rusko* ‘Russia’ to the country’s struggle with the pandemic (*úmrtí-nakažený-nákaza-počet-případ-koronavirus-pandemie* ‘death-infected (adj)-infection-number-case-coronavirus-pandemic’), ANTS associates ‘Russia’ minimally with COVID: *koronavirus* is the only one out of 152 associated KWs.

Instead, ‘Russia’, co-occurring with (*EU-NATO-Čína-Německo-USA-OSN*, ‘EU-NATO-China-Germany-USA-UN’), is covered by ANTS as a major international player. It is also associated with the former Soviet satellites, states in former Soviet spheres of influence and with the USSR (*sovětský-Ukrajina-Bělorusko-ČR-Polsko* ‘Soviet (adj)-Ukraine-Belarus-Czechia-Poland’). We have seen in §4.3.4 that ‘Russia’ occurs in contexts of dispute and conflict. The AA for ‘Russia’ is consistent with this observation; here ‘Russia’ occurs not only in conjunction with Turkey and Greece, but also in other contexts of conflict and war (*Sýrie-Turecko-Řecko-Kypr-Donbas-Krym* ‘Syria-Turkey-Greece-Cyprus-Donbas-Crimea’). It is noteworthy that its MS counterpart is associated with only three countries—Belarus, Brazil, and the US. Connections to military actions and security are numerous: (e.g., *válka-voják-vojenský-základna-vojsko-raketa-obrana-bojový-tajný-bezpečnost* ‘war-soldier-military (adj)-[military] base-army unit-missile-defense-combat (adj)-security’); the MS counterpart lists none of these associated KWs. The associated KWs in the ANTS-AA, which are loosely connected to the concept of power and influence in the international arena, are unique to ANTS.

- (26) *Putinem navrhnutý Summit pěti zemí—stálých členů Rady bezpečnosti OSN (Rusko, Čína, USA, Francie, Velká Británie) bude hrát důležitou roli při hledání společných odpovědí na moderní výzvy a hrozby a prokáže společný závazek duchu aliance...* (Rusko: bezpečnost-OSN) (id: 2020-06-21_61_188)

'A summit of five countries proposed by Putin—the permanent members of the **UN Security Council** (**Russia**, China, the US, France, the UK) will play an important role in finding common answers to modern challenges and threats and will demonstrate a shared commitment to the spirit of the alliance...' (Russia: security-UN)

There are other ANTS-specific associated KWs for 'Russia'. These are KWs that reflect the West's confrontation with Russia: economic and diplomatic penalties (*sankce-vyhoštění-západní* 'sanctions-expulsion [of diplomats]-western'), the poisoning of the Russian opposition leader Navalny (*Navalný*), and the possible plan to poison three Czech politicians with ricin. These associations and the text samples below are consistent with the observations above that ANTS produces a persistent narrative of confrontation between the West and Russia. Example (27) concerns a journalist (Ondřej Kundra) who reported that Russian secret agents disguised as diplomats brought a suitcase with ricin and were planning to assassinate three Czech politicians. The Russian embassy denied the authenticity of this story and made Kundra a laughingstock. Example (28) reports the US attempt to obstruct the German-Russian natural gas pipeline project with new sanctions. The expression "events with Navalny" obfuscates what happened to Navalny, thereby not presenting the suspicion that Navalny was poisoned by Russia; the expression "through the lens" also suppresses the cause-effect relationship, i.e., poisoning of Navalny as a reason for the sanctions.

- (27) Title: *Kundra je terčem posměchu kvůli falešné kauze ricin. Od bojovníka proti fake news k dezinformátorovi stačil jeden kufřík* (Rusko: vyhoštění) (id: 2020-06-06_39_55525)

'Kundra is a target of ridicule for his fake ricin case. From fake news fighter to disinformers, one briefcase was enough' (Russia: expulsion)

- (28) *USA se pokoušejí dívat se na "Severní proud" optikou Navalného případu. USA mohou využít událostí s Navalným a zavést nové sankce proti projektu "Severní proud".* (Rusko: Navalný-sankce) (id: 2020-08-31_29_232949)

'The US is trying to look at the "Nord Stream" through the lens of the Navalny case. The US may take advantage of the events with **Navalny** to initiate new **sanctions** against the "Nord Stream" project.' (Russia: Navalny-sanction)

4.3.6. Shared KWs: Summary

Shared KWs are informative when it comes to obtaining evidence for ANTS-unique narratives. Although the KWs are shared by both media classes, ANTS

frame them very differently from MS. The former tends to “flood” the KWs with a much larger number of associated KWs than MS. These large AAs, however, are not arbitrary. Many of the associated KWs recur and weave together a consistent set of narratives: confrontation between the West (USA, EU, Czech Republic as part of EU) and Russia; tendencies to accentuate the negative image of the West in contrast to a positive image of Russia were also observed.

5. Conclusions

This study has explored the discourse properties of the Czech anti-system media and its strategy to “flood the media zone” with various associations. The ANTS texts contained numerous news topics, including the migration crisis in Europe, the Black Lives Matter movement, the Belarus protests, the EU policies, the Czech mainstream public TV, and international conflicts. In ANTS most of them were framed by associations to support specific ideological narratives.

Methodologically, the present analysis was driven by two cognitive principles: contrast and conceptual framing. Prominent words were harvested with Keyword Analysis against the background of a large corpus of journalistic texts; the choice of this reference corpus allowed us to identify features of ANTS that are unusual in the general journalistic register. Market Basket Analysis helped us assess the number and the strength of association among keywords and helped us understand the nature of such associations (represented by keywords) in ANTS, especially when contrasted with MS. Our method is different from both analyses of a few hand-picked texts and analyses of multiple texts on one delimited topic. Individual text samples facilitate our understanding of the nature of associations, but the *spread* of the associations in a large number of texts can only be demonstrated quantitatively.

The results of our investigation have revealed several important features of ANTS discourse. Although ANTS is relatively smaller in size than MS, the ANTS-KWs tend to be associated with relatively more KWs, an indication of “flooding” the media space. These associated KWs, however, hardly occur randomly. Instead, ANTS, using these associations, consistently and repeatedly weaves a small set of recurring narratives, and it does so regardless of what it covers as a news topic: a Cold-War model dividing the world into the West and Russia, consistently focusing on the negative aspects (while being silent about the positive aspects) of the West, thereby implicitly elevating the role of Russia in the international arena. By incrementally connecting different topics to these narratives, ANTS texts create a tacit argumentation for CZexit- and NATO-exit and for reorientation of Czechia towards Russia. These narrative lines are found within the entire ANTS media class in Czech—on both Russia-sponsored and domestic servers.

Acknowledgements

This study was supported by the European Regional Development Fund project “Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World” (reg. no.: CZ.02.1.01/0.0/0.0/16_019/0000734), Norwegian Research Council Grant number 300002 for “THREAT DEFUSER: Mitigating Perceived Threats in Russian and Norwegian Public Discourse,” and Brown University Humanities Research Fund. Many thanks to the anonymous reviewers for the constructive comments and suggestions.

Sources

- Cvrček, Václav, and Pavel Procházka. (2020) *ONLINE: Monitorovací korpus internetové češtiny*. Institute of the Czech National Corpus, Filozofická fakulta Univerzita Karlova. Available at: www.korpus.cz.
- Křen, M., V. Cvrček, T. Čapka, A. Čermáková, M. Hnátková, L. Chlumská, T. Jelínek, D. Kovářiková, V. Petkevič, P. Procházka, H. Skoumalová, M. Škrabal, P. Truneček, P. Vondříčka, and A. J. Zasina. (2019) *SYN*. Version 8. Institute of the Czech National Corpus, Filozofická fakulta Univerzita Karlova. Available at: www.korpus.cz. [Computer software.]

References

- Baker, Paul. (2005) *Public discourses of gay men*. London and New York: Routledge.
- . (2006) *Using corpora in discourse analysis*. London and New York: Continuum.
- Baker, Paul, Costas G. Gabrielatos, and Tony M. McEnery. (2013) *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge: Cambridge University Press.
- Baker, Paul and Tony M. McEnery. (2005) “A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts”. *Journal of language and politics* 4(2): 197–226.
- Cvrček, Václav and Masako Fidler. (2019) “More than keywords: Discourse prominence analysis of the Russian Web Portal Sputnik Czech Republic”. M. Berrocal and A. Salamurović, eds. *Political discourse in Central, Eastern and Balkan Europe*. Amsterdam, Philadelphia: John Benjamins, 93–117.
- . (2021) “By their Associations you will recognize them: Using Market Basket Analysis to probe ‘alternative’ framing of events”. *Corpus Linguistics 2021 Conference*, Limerick, Ireland, July 13–16.
- . (2022) “No Keyword is an island: In search of covert associations”. *Corpora* 17(2): 259–90.

- Eberle, Jakub. (2021) *Spousta lidí nálepkuje jako dezinformace cokoliv, co se jim nelíbí* [A lot of people label whatever they don't like as disinformation]. Available at: <https://www.mediar.cz/spousta-lidi-nalepkuje-jako-dezinformace-cokoliv-co-se-jim-nelibi/>. Last accessed 11 June 2021.
- Fairclough, Norman. (2003) *Analyzing discourse: Textual analysis for social research*. London: Psychology Press.
- Fidler, Masako and Václav Cvrček. (2015) "A data-driven analysis of reader viewpoints: Reconstructing the historical reader using Keyword Analysis". *Journal of Slavic linguistics* 23(2): 197–239. DOI 10.1353/jsl.2015.0018.
- . (2018) "Going beyond 'aboutness': A quantitative analysis of Sputnik Czech Republic". Masako Fidler and Václav Cvrček, eds. *Taming the corpus: From inflection and lexis to interpretation*. Cham: Springer International Publishing, 195–225.
- . (2019) "Keymorph analysis, or how morphosyntax informs discourse". *Corpus linguistics and linguistic theory* 15(1): 39–70. DOI 10.1515/cllt-2016-0073.
- Fillmore, Charles J. (1982) "Frame semantics". Linguistic Society of Korea, ed. *Linguistics in the morning calm*. Seoul, South Korea: Hanshin Publishing Company, 111–37.
- Gabrielatos, Costas G. and Anna Marchi. (2012) "Keyness: Appropriate metrics and practical issues". Paper presented at CADS International Conference, Corpus-Assisted Discourse Studies: More Than the Sum of Discourse Analysis and Computing?, Bologna, Italy, 13–14 September 2012.
- Han, Jiawei, Micheline Kamber, and Jian Pei. (2011) *Data mining: Concepts and techniques*. 3rd ed. Burlington, MA: Morgan Kaufmann.
- Hofland, Knut and Stig Johansson. (1982) *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities.
- Illing, Sean. (2020) "'Flood the zone with shit': How misinformation overwhelmed our democracy". Vox (16 January 2020). Available at: <https://www.vox.com/policy-and-politics/2020/1/16/20991816/impeachment-trump-bannon-misinformation>. Last accessed 19 June 2021.
- Information Resources Management Association. (2014) *Computational linguistics: Concepts, methodologies, tools, and applications*. 1st ed. Hershey, PA: IGI Global.
- Janda, Laura A., Masako Fidler, Václav Cvrček, and Anna Obukhova. (2023) "The case for case in Putin's speeches". *Russian linguistics* 47(1): 15–40. Available at: <https://link.springer.com/article/10.1007/s11185-022-09269-2>
- Lakoff, George. (2014) *Don't think of an elephant!* 2nd ed. White River Junction, Vermont: Chelsea Green Publishing.
- Langacker, Ronald W. (1987) *Foundations of cognitive grammar*. Vol 1, *Theoretical prerequisites*. Stanford, CA: Stanford University Press.
- Manning, Chris and Hinrich Schütze. (1999) *Foundations of statistical natural language processing*. Cambridge, MA: MIT press.

- Oxford English Dictionary*. 3rd ed, s.v. "Flood". Oxford: Oxford University Press. Available at: <https://www.oed.com/>. Last accessed 7 July 2021.
- Popescu, Ioan-Iovitz, Karl-Heinz Best, and Gabriel Altmann. (2007) "On the dynamics of word classes in text". *Glottometrics* 14: 58–71.
- Scott, Mike. (2010) "Problems in investigating keyness, or clearing the undergrowth and marking out trails...". Maria Bondi and Mike Scott, eds. *Keyness in texts*. Amsterdam: John Benjamins, 43–58. Available at: <https://benjamins.com/catalog/scl.41.04sco>.
- Scott, Mike and Chris Tribble. (2006) *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins. DOI 10.1075/scl.22.
- Spahn, Susanne. (2020) *Russian media in Germany*. Available at: <https://www.freiheit.org/sites/default/files/2021-05/russian-media-in-germany.pdf>. Last accessed 11 June 2021.
- Šlerka, Josef. (2018) *Typologie domácích zpravodajských webů* [Typology of domestic media webportals]. Available at: <https://www.nfnz.cz/studie-a-analyzy/typologie-domacich-zpravodajskych-webu/>. Last accessed 10 July 2021.
- van Dijk, Teun A. (2016) "Critical discourse studies: A sociocognitive approach". Ruth Wodak and Michael Meyer, eds. *Methods of critical discourse studies*. 3rd ed. Thousand Oaks, CA: Sage, 62–85.
- Wodak, Ruth and Martin Reisigl. (2016) "The discourse-historical approach (DHA)". Ruth Wodak and Michael Meyer, eds. *Methods of critical discourse studies*. 3rd ed. Thousand Oaks, CA: Sage, 23–61.
- York, Dan. (2010) "Control channel attacks: Fuzzing, DoX, SPIT, and Toll Fraud". Dan York, ed. *Seven deadliest unified communications attacks*. Burlington, MA: Syngress, 93–117. DOI 10.1016/B978-1-59749-547-9.00004-1. Last accessed 8 July 2021.

Masako U. Fidler
Department of Slavic Studies
Brown University
Providence, RI, USA
masako_fidler@brown.edu

Václav Cvrček
Institute of the Czech National Corpus
Charles University
Prague, Czech Republic
vaclav.cvrcek@ff.cuni.cz

The Semantics of Clausal Complementation: Evidence from Polish

Agnieszka Kaleta

Abstract: This paper offers a new approach to post-verbal complement constructions in present-day Polish. The study is couched in the framework of construction grammar theory (cf. Goldberg 1995, 2006; Croft 2001; Diessel 2015). The focus is on four types of complement clauses—the infinitive, gerund, subjunctive, and indicative clauses, which, in keeping with the constructional framework, are taken to represent distinct form-meaning pairings. The main goal of the study is to examine the extent to which these four morphosyntactically different types of complements exhibit differences in meaning and whether there is any semantic patterning in their distribution in present-day Polish. The study employs the method known as collostructional analysis to determine the sets of predicates with which each of the complement constructions is significantly associated and by which it is repelled. The research findings contribute to the semantically based theories of complementation by revealing systematic correspondences between the form and the function of complement clauses, which are modeled in terms of a radial (prototype-based) network of senses. The study provides empirical evidence in support of the thesis that the distribution of (post-verbal) complement constructions is semantically motivated rather than random or arbitrary.

Keywords: complementation, construction grammar, collostructional analysis, gerund, indicative complement, infinitive, subjunctive

1. Introduction

Broadly defined, verbal complements are clauses that function as subject or object arguments of predicates (cf. Givón 2001: 39). The complement clauses that I deal with here are those that appear in post-verbal position and are thus analogous to clausal (nominal) objects. More specifically, I examine the extent to which four morphosyntactically different types of complements—the infinitive, the gerund, indicative clauses, and subjunctive clauses—exhibit differences in meaning and whether there is any semantic patterning in their distribution in present-day Polish.

The study is situated within functional-cognitive approaches to grammar and more specifically within the framework of construction grammar—a fam-

ily of different yet related theories that define grammar as a vast, monostratal repository of constructions, i.e., learned pairings of form with meaning. As defined by Goldberg (2006: 5):

Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency.

Thus, seen from the constructional perspective, all units of a language—from the smallest ones like single words and morphemes to various fixed or semi-fixed expressions, to abstract syntactic patterns, like the passive or the ditransitive—are defined as learned pairings of form with meaning. This view of language presupposes the existence of a lexicon-grammar continuum, which accommodates meanings of different degrees of specificity or generality. While the lexical pole is occupied by meanings that are rich in conceptual detail, the opposite pole accommodates meanings that are maximally general/abstract or schematic. This entails that abstract syntactic templates have meanings of their own, which exist independently of the lexical meanings of the words that happen to fill them. Yet, in order for a lexical item to be “insertable” in a given syntactic slot, there must exist some kind of compatibility between lexical meanings of words, on the one hand, and the schematic meanings of the constructions in which these words occur, on the other (Goldberg 1995). This view has been turned into an effective methodological tool for investigating abstract meanings of schematic constructions. For example, it has been shown in several studies that the English ditransitive construction conveys the general transfer of possession meaning and that this meaning is “recoverable” from the meanings of the verbs that tend to fill the verbal slot in this construction, e.g., *give*, *send*, *bring*, *hand*, *donate* (cf. Goldberg 1995; Stefanowitsch and Gries 2003). Importantly, even if a verb does not carry a transfer meaning in itself, once inserted in the ditransitive construction, it inherits the missing arguments from the constructional template, a process known as semantic coercion (Michaelis 2004; Michaelis and Ruppenhofer 2001). For example, *She baked him a cake* denotes intended transfer, although there is nothing in the semantics of *bake* that would suggest it. This ability of constructional schemas to “override” the meanings of their lexical fillers has been taken as *prima facie* evidence for the existence of constructional (schematic) meanings.

When approached from the constructional perspective, complement clauses such as the infinitive, the gerund, the subjunctive, or indicative clauses are form-meaning pairings, representing the most schematic pole of the lexicon-grammar continuum. These structures have received a considerable amount of attention from functionally and cognitively oriented researchers.

As has been established in the previous research, certain complement types tend to occur with certain semantic types of main-clause predicates. Also, the same predicate may take different complement types depending on its specific senses (cf. Bresnan 1979; Givón 1980, 2001; Noonan 1985; Ransom 1986). This has led to the emergence of the theory according to which the distribution of complement clauses depends on the semantic properties of the main/matrix-clause predicate (see also the early generative accounts). Yet, this theory has come under criticism, mainly for promoting unjustified polysemies of matrix predicates, i.e., positing senses which are unlikely to be represented in the speaker's mental lexicon (cf. Cristofaro 2008; Goldberg 1995). For example, the fact that English perception verbs take both the gerundive complement and the finite complement might suggest the need for positing two different senses for verbs such as *see* or *hear*, i.e., one that designates direct (sensory) perception and another that refers to indirect perception (or inferential reasoning). Hence, the lexical rule theory has been rejected by many scholars, including ones working within the constructional framework, on the grounds that it posits senses that cannot be found in contexts other than the complement constructions themselves (cf. Cristofaro 2008; Goldberg 1995).

Another broad-ranging theory is that individual complement types have meanings of their own and that these meanings contribute to the overall meaning of the sentence. Seen from this perspective, the acceptability of a complementation pattern by a given verb is a consequence of the compatibility between the meanings denoted by the main verb and the meanings associated with the complement pattern itself (cf. Achard 1998; Bolinger 1968; Cristofaro 2008, Dirven 1989; Duffley 2006; Horie 2000; Smith 2008; Wierzbicka 1988). As interesting and revealing as these studies are, they share a common methodological disadvantage in that they rely on introspective methods (i.e., on constructed examples and intuitive judgments of their acceptability), which potentially constrains the scope of semantic generalizations. An alternative approach is the use of corpus-based methods in the study of complement constructions (cf. Egan 2008; Kaleta 2014; Rudanko 2017; Ruohonen and Rudanko 2020; Yoon and Wulff 2016). This approach is much more compatible with the goals and assumptions of construction grammar, which takes a usage-based perspective and defines syntactic constructions as meaningful schemas that emerge as generalizations over actual instances of use (cf. Croft and Cruise 2004; Goldberg 1995, 2006; Langacker 1987, 1991; Tomasello 2003). Thus, when seen from this perspective, studying syntactic constructions entails studying general, schematic representations that motivate and sanction particular "usage events". This, in turn, requires a method that permits a comprehensive detection of all the possible instantiations of a given construction, or at least a representative set of such instantiations, a goal that can hardly be achieved with purely introspective methods of data collection and analysis. Thus, corpus-based methods have by now established themselves as a

standard methodological procedure in constructional frameworks. However, their application to complementation studies has not been as extensive as one would expect, and, importantly, most of the hitherto existing studies are based mainly on English-language data, which means that there is a substantial gap in corpus-based research on complement constructions that needs to be filled with data from less studied languages.

Taking this requirement and the limitations of previous studies as a point of departure, I use a corpus-based method known as collostructional analysis to explore the semantic underpinnings of four main types of Polish post-verbal complement constructions, i.e., the infinitive, the gerund, the subjunctive, and indicative clauses. Following previous research, I assume that matrix verbs hold important clues to the meanings of those constructions. The study relies exclusively on corpus data and, in particular, on exhaustive lists of complement-taking predicates, as extracted from the Polish Web Corpus (cf. §2). With this methodological approach, it was possible to shed new light on the schematic meanings motivating and sanctioning the constructions in question. Significantly, this study brings to light the polysemous structure of the constructions being analyzed, showing that each of them consists of a network of (inter)related senses, radiating from the general schema and centered around the prototypical use.

The paper is structured as follows. Section 2 outlines the method used in this study. Sections 3–6 present the results obtained from applying collostructional analysis to the four complement constructions under consideration (i.e., the infinitive, the gerund, the subjunctive, and indicative clauses). It has been a standard practice in functional/cognitive linguistic studies to contrast different types of complement constructions, as this brings to light the often subtle semantic contrasts exhibited by these structures (cf. Givón 1980; Smith 2008; Wierzbicka 1988). Section 7 offers a detailed discussion of the results presented in the previous sections. The paper closes with some concluding remarks and prospects for future research.

2. Methodology

Collostructional analysis is a collocation-based method which investigates syntagmatic relationships between words and constructions associated with them. Its distinguishing feature is that it ranks words not by raw frequencies, but by their degree of attraction to a construction. Specifically, the method is aimed at studying the semantic properties of schematic constructions by examining how the words filling constructional slots (called *collexemes*) fall into semantic classes (Stefanowitsch and Gries 2003; Gries and Stefanowitsch 2004).

The data for the study have been extracted from the Polish Web 2012 corpus, available via Sketch Engine software (<https://www.sketchengine.eu>; see Kil-

gariff et al. 2014). The relevant patterns have been obtained using the SQL (Structured Query Language) function of the Sketch Engine. Given that Polish Web 2012 is a very large corpus (around 812,818,518 words) and the search constructions are highly conventional and thus very frequent in use, a smaller sub-corpus of approximately 43,226,158 words was created from randomly chosen texts in order to avoid very high token frequencies and also to ensure the possibility of manual inspection (cleaning) of the data where appropriate. The table below presents the token and type frequencies of the constructions being analyzed together with the SQL query codes used for their extraction from the corpus.

Table 1. Data obtained from the Polish Web 2012
(sub-corpus of 43,226,158 words)

Construction	SQL code	Tokens	Types
Verb + infinitive	[tag="V.*"][]{}{0,2}[tag="inf.*"]	463,312	105
Verb + gerund	[tag="V.*"][]{}{0,2}[tag="ger.*"]	16,712	191
Verb + <i>żeby</i> -clause	[tag="V.*"][]{}{0,2}[word="żeby"]	5,856	67
Verb + <i>że</i> -clause	[tag="V.*"][]{}{0,2}[word="że"]	217,859	261

The procedure followed in the collostructional analysis involves calculating the expected frequencies for each of the verbs occurring with a particular construction and comparing them with the corresponding observed frequencies. Next, it is determined whether the deviation observed between these two types of frequencies is statistically significant. The Fisher Exact test is generally recommended for significance testing in collostructional analysis. Yet other statistics are also eligible, if the expected frequencies are higher than 5. Given that this is the case in the present study (i.e., the expected frequencies tend to exceed 5), the z-score test was used to test the statistical significance of the associations between matrix verbs and the complement constructions.

Given that no specific semantic theories concerning Polish complementation are available in the existing literature, the study has an exploratory rather than hypothesis-testing character. That is, the main goal is to uncover the semantic factors that motivate the four complement constructions based on their distributional (collocational) properties. Due to space limitations, the sections to follow present only the top 30 collexemes of each construction. However, given their significance, these predicates tend to be representative of the constructions as whole.

Table 2. 30 most strongly attracted collexemes of the infinitival complement

Collexemes	English	Raw freq.	z-score
móc	can	119,116	1,176.46
musieć	must	54,071	853.58
chcieć/zechcieć	want	51,580	682.31
powinien	should	28,675	628.30
zaczynać/zacząć	begin, start	29,982	521.60
potrafić	know how	15,236	427.04
próbować/spróbować	try, attempt	11,723	360.57
przestawać/przestać	stop	6,933	292.18
trzeba	one should	13,097	278.17
udawać się/udać się	manage	8,121	261.17
starać się/postarać się	endeavor, try	7,342	255.38
umieć	know how, be able to	5,284	241.43
pozwalać/pozwolić	allow, let	8,429	231.90
postanawiać/postanowić	decide	4,478	231.57
zamierzać	intend	3,706	217.59
warto	it is worth	4,802	187.85
kazać/rozkazać	tell, order	2,826	179.66
dawać/dać	give (let, allow)	12,579	173.37
woleć	prefer	2,884	154.05
zdążyć	manage to do on time	2,174	148.50
wystarczyć	suffice, be enough	4,141	148.15
usiłować	endeavor	1,365	135.81
zdołać	succeed, manage	1,366	135.40
pragnąć/zapagnąć	desire	2,383	129.45
lubić/polubić	like	5,015	126.53
pomagać/pomóc	help	4,140	95.46
prosić/poprosić	ask, request	3,345	82.42
decydować/zdecydować	decide	1,287	65.43
uczyć (się)/nauczyć (się)	teach, learn	2,778	64.21
ośmielać się/ośmielić się	dare	349	61.39

3. The Infinitival Complement and Its Collexemes

Table 2 lists 30 collexemes most strongly attracted by the infinitival complement in descending order of their significance. Both imperfective and perfective forms of the verbs have been retrieved where applicable. Note that some verbs do not form aspectual pairs, i.e., they do not have perfective counterparts. In all the tables, the imperfective forms come before the perfective ones.

The verbs combining with the infinitival complement fall into a few distinct, albeit related semantic classes, viz., modal, volition, causation, and aspectual verbs. Table 2 shows that modal verbs have a particular prominence among these collexemes. The single most strongly attracted verb is *móc* 'can', which expresses abilities or possibilities, and also extends to speech acts such as offers and permissions. The second most significant lexeme—*musieć* 'must'—is a modal verb of necessity. Apart from these two verbs, the list of the top collexemes of the infinitival complement features other predicates expressing modal meaning. One of them is *powinien* 'should', which codes different types of obligation. The modal meaning of obligation is also expressed by a range of impersonal (subjectless) verb forms, which occur exclusively in third-person singular, neuter form, e.g., *trzeba* 'one should', *warto* 'it is worth', *wystarczy* 'suffice'. The cluster of modal predicates also includes two nearly synonymous verbs expressing ability to perform an action—*umieć* 'know how' and *potrafić* 'know how, be able to'. Finally, there are two light-verb constructions, one with *mieć* 'have to do' and the other with *dać* 'let', which also convey modal meanings: the former expresses different shades of necessity, while the latter conveys permission.

Another cluster that can be found in Table 2 consists of verbs expressing various desiderative meanings, e.g., wanting, desire, intention, decision to perform an action: *chcieć* 'want', *pragnąć* 'desire', *woleć* 'prefer', *zamierzać* 'intend', *postanawiać* 'decide', and *decydować* 'decide'. The most significant of these is *chcieć*, which comes right after the two most strongly attracted modals (*móc* and *musieć*). All of these verbs refer to future, that is, non-realized actions or events. However, the infinitival construction also combines with some verbs that denote the performance or occurrence of an action rather than hypothetical (future) actions. Here we find verbs denoting attempted action: *próbować* 'try', *usiłować* 'attempt', *starać się* 'try'; verbs of successful action: *udawać się* 'succeed', *zdołać* 'manage', *zdażyć* 'manage', *ośmielać się* 'dare'; and aspectual verbs designating the onset and cessation of an action or event, e.g., *zaczynać* 'begin', *przestać* 'stop'.

Yet another set of infinitive-taking lexemes consists of causative directives such as *kazać* 'tell, order'; *prosić* 'ask, request'; and *pozwalać* 'allow, let'. Most predicates in this category relate to a message directed at the addressee, and their function is to influence the addressee's action. As defined by Searle

(1979), directive speech acts have illocutionary (performative) force, as the main-clause subject tries to get the complement subject to perform an action by means of an utterance. The causing event is usually a verbal rather than a physical action. However, the list of significantly attracted lexemes also features verbs such as *uczyć* 'teach, learn' and *pomagać* 'help', which can express causing events that are physical actions.

Finally, the infinitive forms a significant relationship with verbs of like *lubić* 'like', and the whole construction denotes enjoyment derived by the main clause agent from performing the activity described in the complement clause. Other verbs belonging to this cluster, e.g., *uwielbiać* 'adore', *kochać* 'love', and *nienawidzić* 'hate', either rank much lower or are repelled by the infinitival complement rather than being attracted by it.

4. The Gerund and Its Collexemes

The gerundive construction has turned out to be highly productive in terms of the number of verb types that it felicitously combines with (cf. Table 1). This diversity is not surprising given that the verbs which accept the gerundive complement can also often occur with regular noun phrases. This multiplicity and diversity of verbal collexemes makes it difficult to find regularities in the distribution of the gerundive construction. Yet some coherent sets of verbs can be identified among the most significant matrix verbs, which are listed in Table 3 on the opposite page. The single most strongly attracted collexeme is *umożliwiać* 'enable'. Also, its antonym—*uniemożliwiać* 'disenable'—ranks high, that is, as the fourth most strongly attracted verb. Other collexemes that convey related meanings include *ułatwiać* 'facilitate', *sprzyjać* 'be conducive to', and *służyć* 'serve'. All these verbs tend to appear with non-human subjects, which are construed as instruments that make something possible, or make it possible for someone to do something. Note that the complement agent tends to remain unspecified or has generic reference. Consider examples (1–3)¹:

¹ The following abbreviations are used throughout the paper: 1 = first person, 3 = third person, SG = singular, PL = plural, ACC = accusative, GEN = genitive, DAT = dative, INST = instrumental, PRES = present tense, PST = past tense, INF = infinitive, GER = gerund, IMP = imperative, IMPER = impersonal, REFL = reflexive. As most sentential examples in the paper are relatively long, the specific constructions of direct relevance to the analysis are set off with italics for greater clarity.

Table 3. The 30 most strongly attracted collexemes of the gerundive complement

Collexemes	English	Raw freq.	z-score
umożliwiać/umożliwić	enable	762	306.24
powodować/spowodować	cause	1,025	219.80
odmawiać/odmówić	refuse	639	218.53
uniemożliwiać/uniemożliwić	disable, prevent	318	184.72
zaprzestawać/zaprzestać	cease	214	164.84
ułatwiać/ułatwić	facilitate	329	162.87
wymagać	require	659	159.47
sprzyjać	be conducive to, foster	263	129.02
zapobiegać/zapobiec	prevent	223	124.23
żądać/zażądać	demand	270	117.34
ulegać/ulec	undergo	312	112.61
utrudniać/utrudnić	hamper/impede	195	108.87
zabraniać/zabronić	forbid	216	108.11
unikać/uniknąć	avoid	341	105.39
zakazywać/zakazać	ban, forbid	191	102.38
rozważać/rozważyć	consider	154	86.85
grozić/zagrozić	threaten	263	85.00
skutkować	result in	94	77.46
proponować/zaproponować	propose	301	76.44
służyć/posłużyć	serve	301	76.36
nakazywać/nakazać	order	139	75.65
rozpoczywać/rozpocząć (się)	start, begin	379	75.30
przyspieszać/przyspieszyć	accelerate	124	74.07
znaczyć/oznaczać	entail, mean	508	73.47
zalecać/zalecić	recommend	104	62.75
przewidywać/przewidzieć	envision	191	61.90
planować/zaplanować	plan	240	61.27
postulować	postulate, propose	54	60.01
zapowiadać/zapowiedzieć	announce	177	58.09
zlecać/zlecić	commission, task sb with	61	51.17

- (1) To usługa, która umożliwia odbieranie i
 this tool which enable.3SG.PRES receive.GER.ACC and
wysyłanie e-maili za pośrednictwem urządzeń
 send.GER.ACC emails through means devices
 przenośnych.
 portable
 'This is a tool which enables receiving and sending emails through
 portable devices.'

- (2) Facebook ułatwia nam śledzenie poczynąń
 Facebook facilitate.3SG.PRES us follow.GER.ACC activities
 naszych znajomych.
 our friends
 'Facebook facilitates following our friends' activities.'

- (3) Mantra służy wprowadzaniu umysłu w stan
 Mantra serve.3SG.PRES bring.GER.DAT mind in state
 bezmyślenia.
 thoughtlessness
 'Mantra serves bringing the mind to the state of thoughtlessness.'

A related cluster consists of verbs which express causation: *powodować* 'cause', *skutkować* 'result in', *przyspieszać* 'accelerate'. These verbs also tend to occur with non-human (often processual) subjects, which are nonetheless conceptualized as "causers" rather than instruments, and the complement clause codes the resultant state or situation. This is illustrated in (4–5):

- (4) Zastąpienie godła państwowego godłem uczelni może
 replacing emblem state emblem school can
spowodować obniżenie rangi dyplomów
 cause.INF downgrade.GER.ACC prestige diplomas
 'Replacing the state emblem with the school emblem may cause the
 downgrading of the prestige of the diplomas.'
- (5) Kończę opakowanie termogeniku, który teoretycznie
 finish packet thermogenic which theoretically
przyśpiesza spalanie tłuszczu.
 accelerate.3SG.PRES burn.GER.ACC fat
 'I'm finishing the packet of thermogenic, which is said to accelerate fat
 burning.'

The notion of result or producing a particular effect is also conveyed by predicates such as *wymagać* 'require', *ulegać* 'undergo', *oznaczać* 'entail', and *grozić* 'threaten', as illustrated in the examples below:

- (6) Zmiana hasła wymaga wpisania hasła
change password require.3SG.PRES type.in.GER.GEN password
aktualnie obowiązuje.
currently valid
'The change of password requires entering the current password.'
- (7) Do 2050 populacja Afryki ma ulec podwojeniu.
till 2050 population Africa has undergo.INF double.GER.DAT
'By 2050 the population of Africa is to undergo doubling.'
- (8) Wyrejestrowanie się z portalu oznacza
signing.out REFL from portal mean.3SG.PRES
usunięcie wszystkich usług jakie użytkownik nabył.
remove.GER.ACC all services that user obtained
'Signing out of the portal means removing all the services that the user has obtained.'
- (9) Niedostarczenie skierowania grozi usunięciem
failure.to.deliver referral threaten.3SG.PRES remove.GER.INST
z listy oczekujących
from list waiting
'Failure to deliver a referral might lead to removal from the waiting list.'

Among the top collexemes of the gerund we also find the antonyms of the verbs expressing enablement and causation—*uniemożliwiać* 'disenable', *utrudniać* 'hamper', *zapobiegać* 'prevent'. Other verbs that share this negative orientation are directive speech act verbs that denote the notion of prevention: *zabraniać* 'forbid', *zakazywać* 'forbid, ban'. Note that *zabraniać* is also found on the list of the significant collexemes of the infinitival construction. A closer consideration of the relevant concordance lines shows that the gerundive construction is most common in impersonal contexts, that is, where the identity of the complement clause subject, and often also the identity of the matrix subject, remains unspecified. The infinitive, on the other hand, tends to be found in constructions with expressed and specific subjects. Examples (10) and (11) illustrate this contrast:

- (10) Regulamin serwisu [...] *zabrania* *umieszczania* treści
 regulations service forbid.3SG.PRES publish.GER.GEN content
 pornograficznych [...].
 pornographic
 'Service regulations forbid publishing pornographic content.'
- (11) Rodzice *zabronili* *mi* *gdziekolwiek* *dzwonić* dopóki
 parents forbid.3PL.PST me anywhere phone.INF until
 nie wypełnię swoich obowiązków.
 not fulfil one's.own duties
 'My parents forbade me to phone anywhere until I fulfil my duties.'

A related cluster consists of verbs of negative volition, i.e., *odmawiać* 'refuse' and *unikać* 'avoid'. Yet, unlike verbs of prevention, they accept only the gerundive complement, and the issue of agent specificity or non-specificity appears to be irrelevant in this case. Consider (12) and (13):

- (12) Wdowa Katarzyna Herbert *odmówiła* *przyjęcia*
 widow Katarzyna Herbert refuse.3SG.PST accept.GER.GEN
 odznaczenia.
 decoration
 'Widow Catherina Herbert refused to accept the decoration.'
- (13) Przed snem należy *unikać* *palenia* *papierosów* i
 before sleep should avoid.INF smoke.GER.GEN cigarettes and
picia *alkoholu*.
 drink.GER.GEN alcohol
 'One has to avoid smoking and drinking alcohol before going to sleep.'

Another cluster consists of directives such as *żądać* 'demand', *nakazywać* 'order', *proponować* 'propose', *zalecać* 'recommend', *zlecać* 'commission', and *postulować* 'postulate'. Here the identities of the complement agents tend to remain unspecified, and the constructions convey formal requirements, or recommendations, as illustrated in (14–17):

- (14) Prawnicy *żądali* *wstrzymania* nowego wydania
 lawyers demand.3PL.PST suspend.GER.GEN new issue
 “Baśni” lub *usunięcia* z książki “obscenicznych”
 fairy.tales or remove.GER.GEN from book obscene
 fragmentów.
 fragments

‘Lawyers demanded that the publication of the new edition of “Fairy Tales” be suspended or the obscene fragments be removed from the book.’

- (15) Znow *prawo nakazuje* *publikowanie* oświadczenia
 again law demand.3SG.PRES publish.GER.ACC statement
 majątkowego wójta.
 financial major

‘Again the law demands that the Major’s tax return be published.’

- (16) Ich autorzy *proponują* *wzięcie* udziału w
 their authors propose.3PL.PRES take.GER.ACC part in
 ankiecie i obiecują za jej wypełnienie 75 dolarów.
 questionnaire and promise for its filling 75 dollars

‘Their authors propose taking part in the questionnaire and promise 75 dollars for filling it out.’

- (17) Lekarze *zalecają* *korzystanie* z sauny
 doctors recommend.3PL.PRES use.GER.ACC from sauna
 w celu wzmocnienia organizmu.
 in purpose strengthening body

‘Doctors recommend using the sauna for the purpose of strengthening one’s body.’

The gerundive complement can also be found with aspectual verbs *zaprzątać* ‘cease’ and *rozpoczynać* ‘begin’. They both are rather formal variants of the two other verbs of aspect, i.e., *zaczynać* ‘start’ and *przestawać* ‘stop’, which are significantly attracted to the infinitival complement (cf. §3). These contrasts deserve a study of their own, yet it can be hypothesized at this point that it is again the impersonal aspect that plays a crucial role here: the gerundive constructions tend to de-focus the role of the agents (which are either unexpressed or generic), as a result of which the constructions have a rather formal character. Consider (18) and (19):

- (18) Po moim odejściu ze sztabu *zaprzestano zbierania*
 after my leaving from staff stop.IMPER collect.GER.GEN
 raportów na ten temat.
 reports on this topic
 'After my leaving the staff, they stopped collecting the reports on this subject.'
- (19) Majowie [...] *rozpoczynali liczenie* dni od ważnych
 Mayans start.3PL.PST count.GER.ACC days from important
 wydarzeń [...].
 events
 'Mayans ... started counting the days from important events...'

Finally, let us consider verbs such as *planować* 'plan', *przewidywać* 'envision', *zapowiadać* 'announce', and *rozważać* 'consider'. They are semantically related in that they all refer to the possibility of a future occurrence of an action or event. Like the predicates discussed earlier, they allow de-focusing of the role of the complement agent, who does not have to be strictly coreferential with the main-clause subject and therefore may not be directly responsible for the process described in the complement clause. Consider the following examples:

- (20) Premier Donald Tusk *zapowiedział wycofanie*
 prime.minister Donald Tusk announce.3SG.PST withdraw.GER.ACC
 się Polski z umowy ACTA.
 REFL Poland from agreement ACTA
 'The PM Donald Tusk announced withdrawing Poland from ACTA agreement.'
- (21) Harmonogram prac *przewiduje oddanie* budynku
 schedule work predict.3SG.PRES return.GER.ACC building
 do użytkowania do końca marca 2013 roku
 for use till end March 2013 year
 'The work schedule predicts putting the building into use by the end of March 2013.'

In (20) the Prime Minister is construed as one of the decision-makers rather than someone who will be directly and personally involved in the complement process. By the same token, in (21) the identities of complement agents remain unspecified and also irrelevant. What is at issue here is the possibility

of an event occurring in the future rather than the role of the actors in effecting this occurrence. The two other predicates—*planować* and *rozważać*—also have the effect of de-focusing the agents' involvement in the process described in the complement, albeit they do so in a more subtle way. Consider (22) and (23):

- (22) Założyciel firmy Vook, Brad Inman *planuje*
 founder company Vook Brad Inman plan.3SG.PRES
wydanie w 2010 dwustu dostępnych jedynie
 publish.GER.ACC in 2010 two.hundred available only
w internecie książek.
 in internet books

'The founder of the company Vook, Brad Inman, plans to publish two hundred books in 2010 that will be available exclusively online.'

- (23) Komitet strajkowy *rozważa zakończenie* strajku.
 committee strike consider.3SG.PRES end.GER.ACC strike

'The strike committee is considering ending the strike.'

In (22) the subject referent—as the owner of the publishing company—is not likely to be directly involved in the process of publishing the books. That is, his role is more of a manager or controller of the whole process than its direct participant. In (23), in turn, the strike committee is not necessarily (and strictly) identical with the complement clause agents, i.e., the workers taking part in the strike. Thus, also here the matrix agents have a "supervising" or "controlling" role to play rather than being the ones directly involved in the complement process.

5. *Żeby*-Complement and Its Collexemes

This section presents the results of collostructional analysis, as performed on the *żeby*-complement construction, which corresponds to what English-language literature tends to label as the subjunctive. The Polish subjunctive is a highly polysemous construction, with many different syntactic and semantic functions (cf. Kaleta 2021). Given the focus of this paper, the present analysis is restricted only to the post-verbal forms functioning as object complements. Table 4 on the following page presents the top 30 collexemes of the subjunctive complement.

The single most strongly attracted collexeme is *chcieć* 'want'. Upon inspection of the relevant concordance, it has become obvious that this use is restricted to situations where the matrix subject is non-coreferent with the subject of the complement clause. The same applies to other volition verbs

Table 4. The 30 most strongly attracted collexemes of the subjunctive

Collexemes	English	Raw freq.	z-score
chcieć/zechcieć	want	1,676	211.29
prosić/poprosić	ask	452	119.06
sądzić	judge, suppose	248	112.50
pilnować/dopilnować	see to	134	110.30
wątpić/zwątpić	doubt	79	72.95
ważne	it is important	172	71.65
życzyć/zażyczyć	wish	156	69.73
namawiać/namówić	talk into, persuade	71	61.49
upierać się/uprzeć się	insist on	21	60.77
błagać	beg	57	59.57
żądać/zażądać	demand	81	59.27
modlić się	pray	81	57.74
wyobrażać/wyobrazić (sobie)	imagine	76	55.11
marzyć	dream	72	49.41
możliwe	it is possible	98	47.95
proponować/zaproponować	propose, suggest	110	47.29
naciskać	urge, insist	33	46.01
mówić/powiedzieć	say, tell	510	45.48
pozwalać/pozwolić	allow, let	119	43.30
woleć	prefer	83	41.86
uważać	be careful not to	165	40.39
sprawiać/sprawić	cause, make	76	39.95
przekonywać/przekonać	persuade	84	34.64
starać się/postarać się	try, aim	98	34.42
zależeć	be intent on	81	32.98
zmuszać/zmusić	force, coerce	67	31.45
nalegać	insist	13	29.15
kusić/skusić	tempt	26	28.11
wymagać	demand, require	65	24.86
radzić/doradzić	advise	65	23.99

strongly associated with the *żeby*-complement, such as *życzyć* 'wish' and *woleć* 'prefer'. This, of course, distinguishes the subjunctive uses of these predicates from the corresponding infinitival uses, which require coreferential subjects. Apart from these three prototypical verbs of volition, there are other collexemes that may not be volitional in and of themselves yet acquire volitional meanings when complemented by the *żeby*-clause. They include *modlić się* 'pray', *marzyć* 'dream', *upierać się* 'insist on', *kusić* 'tempt', and *zależeć* 'be intent on'. All of these verbs can be used both in same-subject and non-coreferent constructions. Note that *marzyć* and *modlić się* indicate a strong wish for something that is impossible or unlikely to happen, whether they come in coreferent or non-coreferent constructions. Verbs *upierać się* and *kusić*, on the other hand, denote a strong volition or desire experienced by the subject referent when used in same-subject constructions. However, in non-coreferential constructions, these two verbs acquire causative meaning, denoting pressure imposed on the complement agent by the main agent to get him/her to perform an act. Also, the impersonal construction—*jest ważne* 'it is important'—is a part of this cluster as it combines with the subjunctive clause to express the speaker's wishes or desires concerning a particular outcome. However, it can also express recommendations and suggestions as to a course of action to be followed by others.

These non-coreferent uses of volition verbs appear to be related to another cluster discernible among the significant collexemes listed in Table 4, namely, causation verbs (or manipulation verbs in Givón's 2001 terminology). Most of them are speech act verbs: *prosić* 'ask, request'; *namawiać* 'persuade, talk into'; *brać* 'beg'; *żądać* 'demand'; *naciskać* 'insist'; *proponować* 'propose'; *pozwalać* 'allow'; *przekonywać* 'persuade'; and *radzić* 'advise'. All these verbs express directive acts in which the main-clause subject uses speech to get the complement subject to perform the action described in the complement clause.

It should be noted here that the directives most strongly attracted by the *żeby*-complement are generally weaker than those that take the infinitive. For example, *brać*, which is apparently the weakest of all these predicates in terms of the degree of influence exerted on the complement subject by the main-clause subject, is not to be found among the collexemes of the infinitival construction. The same relates to *namawiać*, which does not accept the infinitival complement, or *radzić*, which is repelled rather than attracted by the infinitival complement. The stronger or more authoritative directives, like *kazać* 'tell, order' and *zabraniać* 'forbid', on the other hand, are either repelled by the subjunctive construction or do not figure at all on the list of its collexemes.

Somewhat puzzling in this context are directives such as *nalegać* 'insist' and *naciskać* 'put pressure on', which, despite expressing rather forceful acts, take the subjunctive, not the infinitival complement. Yet the specific semantics of these verbs provides potential clues to this usage. That is, *naciskać* and *nalegać* imply that the directive is met with some resistance on the part of

the complement agent, who is apparently unwilling to engage in the action described in the complement clause. This, in turn, suggests that the influence exerted over the complement agent by the main agent might not be as strong as in the case of other directives. In other words, given that the complement subject is construed as being capable of acting independently, the role of the main agent as the only or the main decision-maker (or “controller”) is clearly diminished in this case.

Complementation of *żądać* also appears to be puzzling at first sight, especially given that the semantically related *kazać* preferentially co-occurs with the infinitival complement. Yet, it needs to be recognized that these two predicates are distinct in terms of the degree of emotional charge that they encode—while the latter is rather neutral, the former implies some degree of the subject’s emotional involvement, i.e., it conveys a very strong, firm request. Given that emotional charge is only necessary when some opposition is encountered or expected, the subjunctive complement appears to have here the same motivation as in the case of the two insistence verbs (i.e., *nalegać*, *naciskać*). That is, the implication of the resistance on the part of the complement agent coincides with a greater likelihood of the main agent’s authority being challenged in one way or another (cf. Givón 1980: 368).

Note that the notion of insistence or imposing one’s will on another (usually resistant) agent in order to get them to perform an act can also be conveyed by manipulation predicates, which do not necessarily involve speech. The most significant of them are *pilnować* ‘see to’, *zmuszać* ‘force’, and *wymagać* ‘demand, require’. Two other verbs which appear to cluster with these predicates are causatives *starać się* ‘try’ and *sprawić* ‘cause’. They are, however, distinct in that they profile the result achieved by means of the action performed by the main agent rather than acts of verbal or other coercion. This is illustrated with the following examples:

- (24) *Staram się, żeby moje lekcje były naprawdę*
 try.1SG.PRES REFL so.that my lessons be.3PL.PST really
 ciekawe.
 interesting

‘I try to make my lessons really interesting.’

- (25) *Ty wiesz jak sprawić żeby kobieta czuła się*
 you know how make.INF so.that woman feel.3SG.PST REFL
 ważna.
 important

‘You know how to make a woman feel important.’

Another distinct cluster that can be distinguished among the collexemes of the subjunctive construction consists of predicates describing mental states: *sądzić* 'think', *uważać* 'think, believe', *wątpić* 'doubt', *wyobrażać sobie* 'imagine', *możliwe* 'it is possible'. The main function of these constructions is to express the subject's opinions concerning the truth of the complement proposition. Note that all these verbs, except *wątpić*, must be negated in order to be compatible with the subjunctive complement. What all these constructions have in common is that the speaker does not present the facts but merely evaluates them, expressing a certain degree of uncertainty or disbelief concerning the veracity of the proposition being presented in the complement clause. Consider (26) and (27) as examples:

- (26) *Nie wyobrażam sobie, żebyśmy kiedykolwiek mieli*
 not imagine.1SG.PRES REFL so.that ever have.3PL.PST
się rozstać.
 REFL split.up
 'I cannot imagine that we will ever split up.'

- (27) *Nie uważam, żeby był jakimś strasznym*
 not think.1SG.PRES so.that be.3SG.PST some terrible
alkoholikiem.
 alcoholic
 'I don't think that he is a heavy drinker.'

6. *Że* 'That'-Clauses and Their Collexemes

Finite *że*-clauses have been found to combine with as many as 261 different verb types, out of which 207 have a positive association with this complement type. Despite the high productivity of this construction, its collexemes form a rather consistent group, comprising a few distinct, yet related, sub-clusters centered around the most strongly attracted verbs. Table 5 on the following page lists the top 30 collexemes. The most common use of *że*-clauses is to report what somebody said. This use is represented by the most strongly attracted *twierdzić/stwierdzać* 'claim', but also by a range of other verbs of speech, which include *mówić* 'say, tell', *przyznawać* 'admit', *dodawać* 'add', *podkreślać* 'emphasize', *oświadczać* 'announce', *wmawiać* 'convince', *sugerować* 'suggest', and *przekonywać* 'convince, persuade'. Note that some of these verbs can also occur with the subjunctive complement. This, however, most often entails a change in meaning. Let us consider, for example, *sugerować*, as exemplified below:

Table 5. The 30 most strongly attracted collexemes of *że*-construction

Collexemes	English	Raw freq.	z-score
twierdzić/stwierdzać/stwierdzić	claim	11,322	474.47
uważać	believe, think	9,332	391.76
mówić/powiedzieć	say, tell	7,574	362.59
okazywać się/okazać się	appear, turn out	23,479	361.76
wiedzieć/dowiedzieć się	know, learn, find out	19,663	361.61
sprawiać/sprawić	cause, make	6,066	343.30
myśleć/pomyśleć	think	6,626	288.28
przyznawać/przyznać	admit	5,082	255.91
znaczyć/oznaczać	mean, entail	6,250	252.72
sądzić	think	2,958	217.51
wierzyć/uwierzyć	believe	5,234	209.17
wydawać się	seem	969	152.12
przypuszczać	suppose	3,240	151.78
pamiętać/zapamiętać	remember	2,280	148.55
zauważać/zauważyć	notice	3,932	148.30
podejrzewać	suspect	1,148	147.64
udowadniać/udowodnić	prove	1,465	144.16
dodawać/dodać	add, mention	3,215	140.16
podkreślać/podkreślić	emphasize	1,810	138.72
cieszyć się/ucieszyć się	be happy	2,823	138.48
oświadczać/oświadczyć	announce	1,160	134.93
wmawiać/wmówić	convince	900	131.99
pokazywać/pokazać	show	1,136	124.66
sugerować/zasugerować	suggest, imply	3,166	124.40
obawiać się	fear	1,189	123.09
przekonywać/przekonać	convince, persuade	1,870	121.62
wynikać/wyniknąć	it follows that	1,986	120.79
rozumieć/zrozumieć	understand	3,634	119.21
powodować/spowodować	cause	2,200	118.74
uznawać/uznać	acknowledge	2,269	114.88

- (28) Naukowcy *sugerują*, *że* zakupoholizm często
 scientists suggest.3PL.PRES that shopaholism often
 związany jest z syndromem wyprzedazy.
 connected is with syndrome sales
 'Scientists suggest that shopaholism is often related to the sales
 syndrome.'

- (29) Minister obrony Amir Peretz *zasugerował*, *żeby* Izrael
 minister defense Amir Peretz suggest.3SG.PST so.that Israel
 rozpoczął negocjacje z Syrią.
 start.3SG.PST negotiations with Syria
 'The defense minister Amir Peretz suggested that Israel should start
 negotiations with Syria.'

The contrast between these two uses is rather obvious: (29) conveys a mild directive, while (28) expresses the subject's standpoint on a particular issue. Semantic shifts of this type clearly show that complement clauses are constructions in their own right, that is, they make their own contributions to the meaning of the whole utterance.

There are approximately 90 different speech verbs among the collexemes of the indicative complement, which accounts for nearly half of all the verbs attracted by this complement. All these verbs express acts of conveying information or knowledge through speech, although they do so in different ways and for different purposes. Among these verbs one can find verbs reporting pure utterance acts, e.g., *informować* 'inform', *oznajmiać* 'state', *zapowiadać* 'announce'; various assertives, e.g., *zapewniać* 'assure', *argumentować* 'argue', *wyjaśniać* 'explain'; or verbs which convey commissive acts such as promises or threats, e.g., *obietcywać* 'promise', *przyrzekać* 'swear', *ostrzegać* 'warn', and *grozić* 'threaten'.

Another common use of *że*-clauses is to report people's mental states and processes. The predicates that belong to this category fall into two distinct, yet related, sets. One of them is that of having or acquiring knowledge in the broad sense, including verbs of memory, learning, and perception: *wiedzieć/do-wiadywać się* 'know, find out', *rozumieć* 'understand', *zauważyć* 'notice'. Another set comprises predicates referring to opinions and beliefs, e.g., *myśleć* 'think'; *sądzić* 'think', *uważać* 'think, believe', *wierzyć* 'believe', *przypuszczać* 'suppose', and *podejrzewać* 'suspect'. Note that mental verbs with *że*-clauses are an important device used to express stance. They encode different degrees of the subject's commitment to the truth of the proposition expressed by the embedded clause. The claims they make vary from full commitment to partial commitment to denial. For example, verbs such as *sądzić* or *uważać* convey a

sense of possibility combined with uncertainty, while verbs such as *wiedzieć* or *dowiedzieć się* convey a definite sense of certainty.

The two verb classes, as presented above—i.e., verbs of speech, on the one hand, and verbs of mental states, on the other—appear to be related in that speech is typically a manifestation of states or processes of thought. Simply put, what one says is what one knows or thinks to be the case. Seen from this perspective, knowing something is conceptually more basic than saying it in that ‘saying’ presupposes ‘knowing’ something, not vice versa (cf. Wierzbicka 1988: 133).

Another cluster consists of verbs expressing emotional states, which, like all the other *że*-taking verbs discussed thus far, are verbs of mental processes rather than action. The top 30 collexemes include only two verbs that can be classified as verbs of emotion—*cieszyć się* ‘be happy’ and *obawiać się* ‘fear’—yet more verbs of this type can be found lower on the list. The emotional states denoted by these verbs can be regarded as having an epistemic component in that they arise as the experiencer’s (mental) response to certain situations rather than being purely bodily reactions to a physical stimulus. To illustrate, in example (30) below, the feeling of joy experienced by the subject referent is a result of their being aware (having knowledge) of the situation described in the complement clause (‘I know that you like it, and this makes me feel happy’). By the same token, (31) refers to predicting a future occurrence, which arguably is an epistemic process involving one’s knowledge of the way things are or will be, usually derived from earlier experience (‘I know that she might not like it, and this causes my fear’).

- (30) *Cieszę się, że ci się podoba.*
 be.happy.1SG.PRES REFL that you REFL like
 ‘I am happy that you like it.’

- (31) *Obawiam się, że może się jej to nie spodobać.*
 fear.1SG.PRES REFL that may REFL her it not like
 ‘I fear that she might not like it.’

Yet another cluster comprises verbs such as *udowadniać* ‘prove’, *pokazywać* ‘show’, *okazywać się* ‘turn out’, *wynikać* ‘follow’, and *oznaczać* ‘entail’. All these predicates communicate some knowledge, though they usually do so in ways that do not involve speech. Given that most of these predicates indicate the source of the knowledge, they may be considered to have evidential function. Note that they often indicate the degree of certainty associated with the reported information. For example, *pokazywać* ‘show’, *udowadniać* ‘prove’, and *dowieść* ‘prove’ mark a high degree of certainty, while a lesser degree of cer-

tainty is expressed by, for example, *sugerować* 'suggest' and *wskazywać* 'indicate'. Example (32) illustrates these uses:

- (32) Przegląd literatury naukowej dotyczącej agresji
 review literature scientific about aggression
pokazuje, że można ją definiować, opisywać i
 show.3SG.PRES that one.can it define describe and
wyjaśniać na wiele sposobów.
 explain in many ways
 'The review of the scientific literature shows that aggression may be defined, described, and explained in many different ways.'

Thus far, we have seen that the indicative complement consistently combines with verbs denoting different types of knowledge. Yet, this uniformity is clearly disturbed by *sprawiać* 'cause, make happen' and *powodować* 'cause'—periphrastic causatives, which encode indirect causation. Consider the example in (33):

- (33) Twój urok osobisty i pogoda ducha *sprawiają,*
 your charm personal and cheerfulness spirit cause.3PL.PRES
że inni czują się przy tobie dobrze i bezpiecznie.
 that others feel REFL with you good and safe
 'Your personal charm and optimism make others feel good and safe in your company.'

In this case, the *że*-complement codes a resultant state or situation brought about by the causing event, as described in the main clause. As noted above, the causal link between these two events is indirect in that the causer does not act immediately and physically on the causee. Significantly, the interpretation of (33) is necessarily based on inference, i.e., cause-effect reasoning, which is arguably an epistemic activity. Note that if the focus is on intentional action ('I want you do something') rather than the resultant state and cause-effect relationship, *sprawiać* combines with the subjunctive complement, not the indicative one. This is illustrated in (34) below.

- (34) *Spraw, żebym poczuła się szczęśliwa.*
 cause.IMP so.that feel.1SG.PST REFL happy
 'Make me feel happy.'

Thus, in cases like this, it is the role of the complement construction to indicate which construal is more relevant. This, in turn, clearly shows that com-

plement clauses have meanings of their own, which are independent of (albeit compatible with) the lexical meanings of matrix predicates.

7. Discussion

The analyses conducted in this study shed new light on the distributional properties of the four main types of complement constructions and hence provide new insights into their semantic structure. A very general distinction can be made between two types of meanings, which represent two general domains, i.e., the deontic (effective) and the epistemic one (cf. Langacker 2010). The former generally refers to actions and events that involve causation of events and thus have an effect on the outside world. The latter involves mental activity, which does not have such an effect. In other words, epistemic predicates describe events, “which can be assessed for validity, but not caused in the way that any causal theory of action will endorse” (Langacker 2010: 166). As we have seen, the deontic (effective) domain is typically represented by non-finite constructions, whereas the epistemic domain tends to be expressed with the finite *że*-complements. The *żeby*-complement, on the other hand, has both deontic and epistemic uses, which means that the Polish subjunctive is a truly linking mood, providing a connection between these two broad domains.

As has been shown in sections 3 and 4, the infinitival and gerundive complements construe the effective domain in two different ways. The collostructional analysis of the infinitival construction has revealed a few distinct, yet related clusters of senses, which evidences the polysemous character of this construction. One of these meanings is the notion of potentiality for an action, as conveyed by the high-ranking modals, with *móc* ‘can’ at the very top of the list. In cognitive linguistic research, modal verbs are described as force-dynamic categories, which involve some conception of potency, that is, “a physical or mental force that, when unleashed, tends to bring about an occurrence of that process” (Langacker 1991: 270). Apart from the modal and semi-modal verbs, the notion of “potency” can be traced in the second major cluster of senses, i.e., the one referring to the concept of volitionality (‘wanting something to happen’), which appears to constitute the semantic prototype of the infinitival construction. As has been established, the volition-related notions such as wishes, desires, hopes, plans, and intentions figure prominently in the semantic network of the construction in question. This means that the actions expressed in the infinitival clause tend to be non-realized, i.e., located in the future with respect to the time of the main-clause event. The other uses appear to represent elaborations or extensions from this prototype, as indicated by their lower frequency in the corpus data. Hence, the predicates designating successfully completed, usually effortful actions (e.g., *zdołać* ‘succeed, manage’, *udać się* ‘manage’) are further examples of volitive behavior.

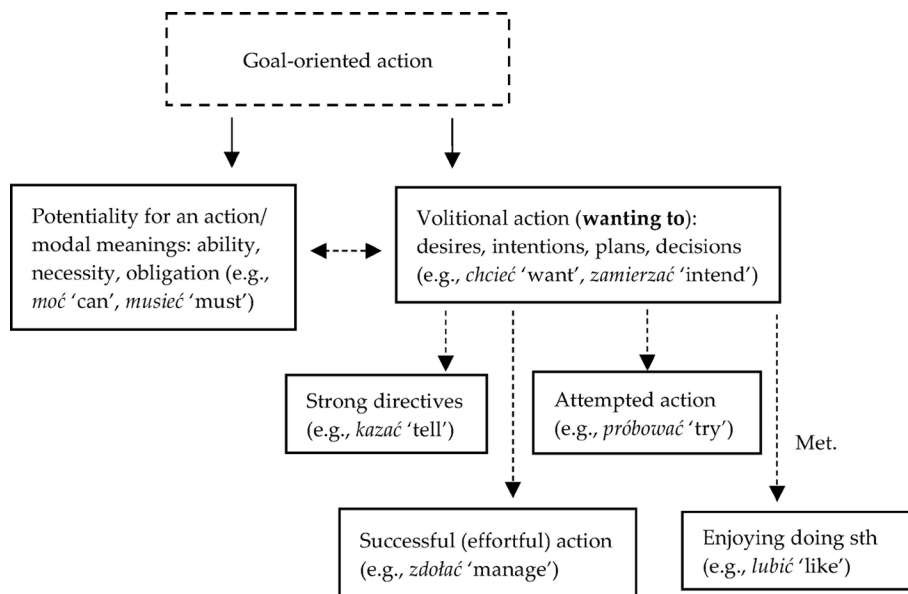


Figure 1. The semantic structure of verb + infinitive construction

However, they express posteriority or subsequence rather than futurity in the strict sense of the word. Also, the strong directives (e.g., *kazać* 'tell', *rozkazać* 'order') convey desiderative meaning, albeit in a different form. While in the basic uses we have to do with self-induced actions (X wants to do Z), in the directive constructions the matrix agent expresses volition aimed at getting someone else to perform an act (X wants Y to do Z).

Finally, there are constructions with verbs of liking (*lubić* 'like'), which, at first sight, might appear to be unrelated to the volitional uses. Yet, as has been convincingly argued by de Smet and Cuyckens (2005) in their study of English complement constructions, the notion of desire and that of enjoyment are closely intertwined in experience in that the things we want are very often the things that we like or love, or vice versa. Hence, it can be postulated that the 'liking' sense is metonymically related to the desiderative meanings.

The diagram in Figure 1 depicts the polysemous structure of the infinitival construction. The dashed box indicates a schema which is represented in all the uses of the infinitival construction, without necessarily "surfacing". That is, this "overarching" schema appears to be that of a very general notion of a goal-oriented action. It is instantiated by the two major (interrelated) clusters of senses, i.e., modal and volitional meanings, both of which represent force-dynamic categories in the sense of Talmy 1988. The other volition-related meanings have been represented as extensions or elaborations of the

prototypical concept of volitionality (the dashed arrows represent extensions, while the solid ones denote instantiations).

Turning now to the gerund, we can see a significant shift in perspective in that the subject's volitionality or the notion of goal-orientedness have no role to play here. On the contrary, the gerundive constructions show a strong tendency to leave the identities of complement agents unspecified or implicit. This, together with the rather diverse range of meanings it conveys (e.g., enabling or causing events, producing particular effects, expressing requirements/suggestions, planning/predicting future events), leads to the conclusion that the gerund has no special semantic significance, apart from highlighting the very general notion of occurrence of an action or event. Another major function of the gerundive complement is to denote the non-occurrence of actions or events, as seen in the uses that refer to preventing events and negative volition. Also, here the identities of the agents may and often do remain unspecified and the constructions tend to have a rather formal and impersonal character. Yet, even if the subject is definite or specific, the mere fact that the gerund codes the non-occurrence of an action or event (refusing, avoiding doing something) entails that the subject referent is not to be construed as an active participant of the complement scene. This conclusion extends to the less obvious uses in which the subject of the main clause is not strictly coreferential with the complement agents and hence not directly or personally involved in the action being described. What seems to follow from all these considerations is that the main function of the gerundive construction is that of de-focusing the role of agents in bringing about a given state of affairs and turning the spotlight on the occurrence or non-occurrence of the complement event as such. The diagram in Figure 2 schematizes the semantic structure of the verb + gerund construction.

The subjunctive is used to structure the effective (deontic) domain, along with the two non-finite complements, as discussed in the previous sections. Like the infinitival construction, it tends to express desiderative meanings (wanting to do sth), the main difference being that the subjunctive preferentially selects constructions with non-coreferent subjects, whereas the infinitive tends to be constrained by the sameness-of-subjects restriction. The exception to this are predicates such as *marzyć* 'dream' or *modlić się* 'pray', which can be complemented by the subjunctive despite having co-referent subjects. Yet, this distribution is not necessarily random or contradictory. In fact, it appears to coincide with the patterns observed cross-linguistically and discussed in some detail in typological literature on clausal complementation. For example, Givón (1980, 2001) argues that there exists an iconic relation between the degree of semantic integration between the main-clause event and the complement event, on the one hand, and the degree of morphosyntactic integration between the matrix and the complement clause, on the other. The relation between the semantic and syntactic structure is such that "the stronger is the

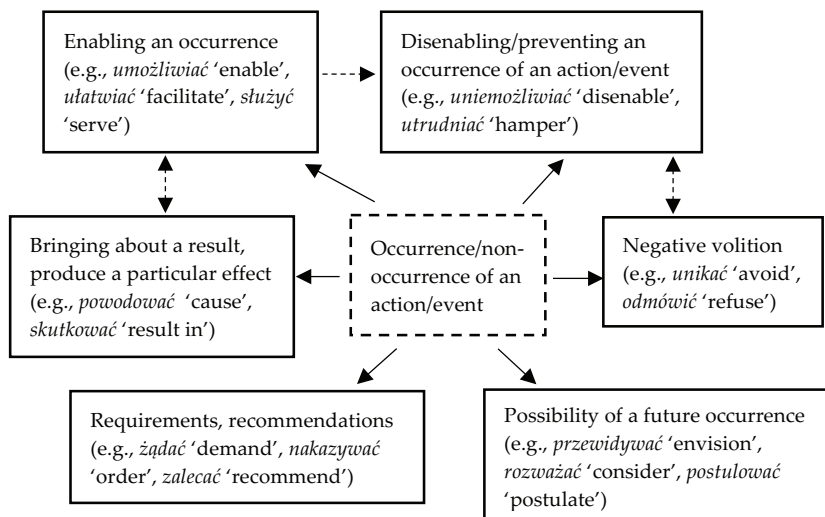


Figure 2. The semantic structure of verb + gerund construction

semantic bond between two events, the more extensive will be the syntactic integration of the clauses into a single though complex clause" (Givón 1980: 337). Within Givón's theory, the binding strength of complement-taking predicates is determined, among others, by the degree of control (causative influence) that the main-clause agent can rightfully expect to exert over the complement agent. Seen from this perspective, the non-coreferent constructions suggest less influence than same-subject constructions, for the simple reason that we tend to have more control over our own actions than over the actions of others. Consequently, the former tend to exhibit less syntactic integration than the latter. Same-subject volitives, i.e., *marzyć żeby* and *modlić się żeby*, are no exceptions in the light of this theory, given that they both indicate that the complement proposition is rather unrealistic and hence beyond the subject's direct control. The same idea seems to extend to directive speech act verbs. As we have seen, the subjunctive tends to occur with weaker directives—the ones with less deontic force and hence less causal influence on the complement agent (cf. *prosić* 'ask, request' and *błagać* 'beg'). Emotionally charged or insistent directives or causatives (e.g., *nalegać* 'insist', *zmuszać* 'force') appear to represent the other side of the same coin in that emotional charge is needed only when some opposition is expected and when one's authority is at stake. Hence, the general pattern that emerges from these data is that the subjunctive codes weak manipulation, in the sense that the causative influence exerted by the main-clause subject over the complement event is restricted in one way or another.

The subjunctive construction extends beyond the deontic (effective) domain of willful action and causation to the epistemic domain, conveying notions such as disbelief or uncertainty. When considered at a higher level of abstraction, these two uses, i.e., weak causation and uncertainty, appear to be (metonymically) related in that the degree of causal influence that one has over an event translates itself into the degree of certainty with which one can predict the occurrence of this event. Given this duality of its semantic structure, the Polish subjunctive proves to be a truly linking mood, allowing a smooth transition between the deontic domain of actions and/or events and the domain of knowledge, as represented by the finite *że*-constructions. The diagram in Figure 3 below presents the conceptual structure of the subjunctive.

Finally, let us consider the semantic make-up of the indicative complement. As has been seen, *że*-clauses are most strongly associated with verbs of speech and verbs of knowledge. Given that speech is secondary to what one knows, the *że*-construction has been considered as having primarily epistemic function. Of course, knowledge is understood here broadly as any form of mental activity involving the storage, acquisition, or transfer of knowledge (including verbs of memory, perception, opinion/belief, and emotional predicates). The only “exceptions” to this overarching tendency are the causation predicates *sprawić/sprawiać* ‘cause, make happen’ and *powodować/spowodować* ‘cause’, which cross the boundary of the epistemic domain proper and extend into the domain of “cause-effect” relationships, which is a special case of ‘knowing’ something, as discussed in §6. This is represented in Figure 4.

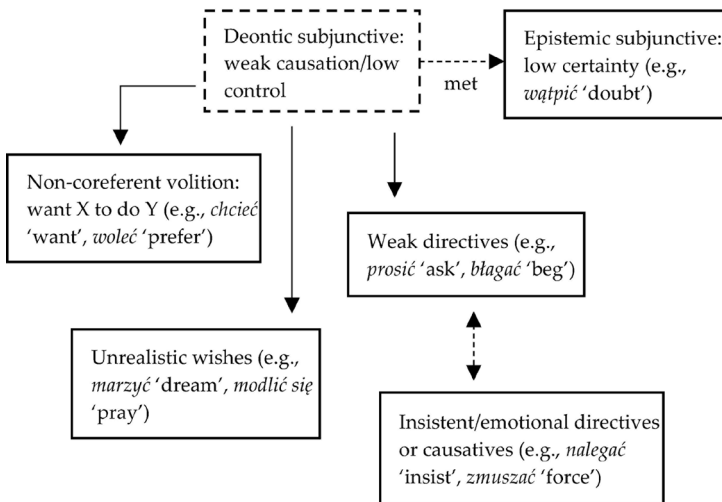


Figure 3. Semantic structure of verb + *żeby*-construction

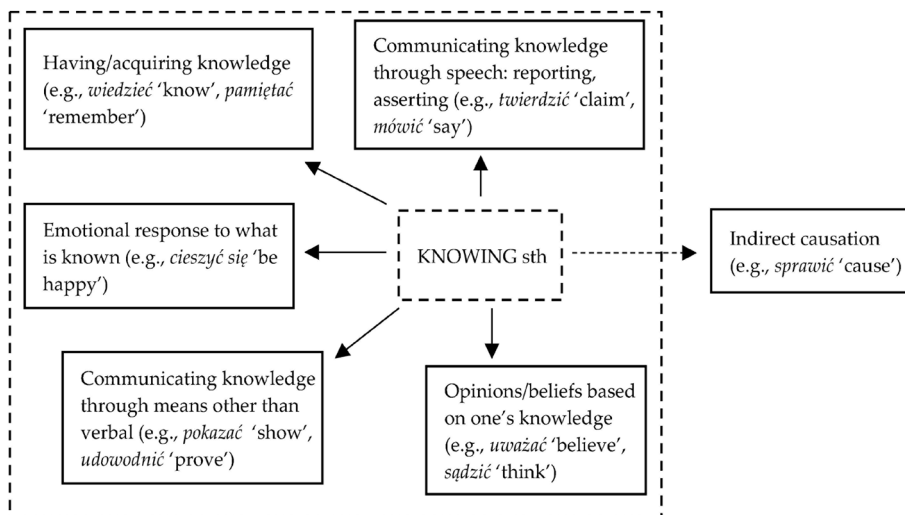


Figure 4. Conceptual structure of verb + *że*-construction

Apart from verbs that take either a non-finite or finite complement type, as discussed thus far, there are also predicates that accept both types. As is argued by Langacker (2010: 180), “[w]hen the same predicate occurs with different complements [...], it has subtly different values which either permit or reflect this usage”. With quite a few predicates, these alternations produce clear semantic contrasts of the sort predicted by our analysis. Compare, for example, (35) and (36):

- (35) *Przekonałem ich, żeby wyjechali.*
 persuade.1SG.PST them so.that left.3PL.PST
 ‘I persuaded them to leave.’

- (36) *Przekonałem ich, że mam rację.*
 persuade.1SG.PST them that have right
 ‘I convinced them that I am right.’

In (35) the result of persuasion is intention of the complement subject to perform an act at some time in the future, whereas in (36), the result is a belief that something is the case. However, apart from such clear semantic contrasts, a number of more subtle semantic shifts can be observed with regard to verbs taking both non-finite and finite complements.

For example, it has been seen that some directive speech act verbs such as *prosić* ‘ask, request’ and *pozwolić* ‘allow’ combine with both infinitive and

subjunctive complements, albeit with a different strength (cf. Tables 2 and 4). However, it is not only the strength of attraction between the matrix verb and the complement that distinguishes these constructions. To illustrate, even a cursory consideration of the concordance lines for *prosić* reveals that *prosić* + infinitive favors predicates in the first-person singular, present tense, active voice. The corresponding subjunctive construction (*prosić* + *żeby*), on the other hand, tends to express third-person reports. Compare (37) and (38):

- (37) *Proszę* *powiedzieć* żonie, że zazdroszczę jej
 request.1SG.PRES tell.INF wife that envy her
 takiego faceta u boku.
 such guy by side

'Please tell your wife that I envy her such a guy by her side.'

- (38) *Mama* *poprosiła* ją, *żeby* *poszła* do psychologa.
 mother ask.3SG.PST her so.that go.3SG.PST to psychologist

'Mother asked her to go to the psychologist.'

The verbs accepting two or more complement types arguably deserve a study of their own but could not be discussed here due to space constraints.

8. Concluding Remarks

A central assumption of functional-cognitive linguistics is that linguistic knowledge consists of the knowledge of constructions, i.e., symbolic units that connect form with meaning. Research in construction grammar has brought to light hundreds of different form-meaning pairings, elucidating the ways in which they function in language and in the minds of language users. The present paper constitutes a contribution to this large body of research by focusing on constructions which have been hitherto relatively poorly understood. Although constructionally oriented research takes it for granted that abstract syntactic templates such as the infinitival or gerundive complement are meaningful in and of themselves, the answers to the question of the semantic import of these constructions have not been clearly spelled out in the previous literature. Collostructional analysis has made it possible to remedy this unfortunate situation to a certain extent by providing an insight into the distributional/ semantic contrasts exhibited by the four main types of complement constructions, as used in present-day Polish. As has been demonstrated, the infinitival, gerundive, and two finite complements (the subjunctive and indicative clauses) represent constructions in their own right, that is, distinct form-meaning pairings, which entails that their distribution is semantically motivated rather than arbitrary. The type and amount of semantic patterning

that has been brought to light in this study appears to provide ample evidence for this thesis. In particular, collostructional analysis has proved fruitful in revealing the polysemous/radial nature of the constructional meanings. Given the high degree of schematicity exhibited by the meanings coded by syntactic categories (as compared to lexical meanings), it should be clear that such an analysis would not be feasible with purely introspective methods.

Complementation is a vast and complex area of study whose full treatment goes beyond the scope of a single paper. As I mentioned in the previous section, chief among the issues that warrant further examination are alternating complement constructions. When seen from a cognitive linguistic perspective, constructions are hierarchical structures, which can be characterized at progressively more specific levels of detail (cf. Langacker 1999). This paper has offered an insight into what can be referred to as “macro-constructions”, i.e., the network representations associated with the schematic syntactic pattern [verb + complement]. The next essential step is to look at “micro-constructions”—the specific instantiations of these general structures, including the verbs that accept two (or more) complements, where a change in complement type produces more or less subtle differences in meaning. It should be clear that capturing those subtle semantic shifts requires a different methodological approach, i.e., one that takes into consideration a variety of specific morphosyntactic and semantic features that can potentially differentiate between two (or more), usually nearly synonymous constructions. I believe that the present study has paved the way for such fine-grained studies by providing a set of hypotheses that can be tested empirically. This approach is consistent with Dirk Geeraerts’s (2010: 73) proposal that corpus-based research is best approached as a cyclical, helix-like process “in which several rounds of data gathering, testing of hypothesis, and interpretation of the results follow each other”.

Finally, there is one caveat to the present analysis. Namely, the schematic semantic representations discussed in this paper cannot be expected to have the predictive or constraining power of the sort assumed in more traditional approaches to semantic analysis. A functional-cognitive linguistic solution to this problem is that general schemas coexist in the minds of language users with a large body of item-specific knowledge, which is stored redundantly (cf. usage-based theory). Hence, while general schemas do have an explanatory and sanctioning role to play, they do not need to be fully predictive, as there are other points of reference that the speakers of a language have at their disposal. As I have indicated throughout this paper, semantic prototypes of constructions provide such points of reference in that they are a crucial and rich source of information about constructional semantics. Thus, establishing which of the semantic representations associated with a polysemous construction is most central or the prototypical one remains an important goal of research in cognitive semantics. Yet, one must also recognize the inher-

ent difficulties involved in accurately defining constructional prototypes (cf. Lemmens 2015; Taylor 2019). While the present analysis has hinted at such representations based on the quantitative data, no systemic, empirical validation of these observations has been undertaken due to space limitations. Hence, another important empirical task is to verify the assumptions regarding the constructional prototypes on the basis of other sources of information, as recommended in cognitive linguistic research, i.e., diachronic data, language acquisition data, or lexicographical resources (cf. converging evidence hypothesis, Langacker 1999). The analysis of the diachronic development of constructional meanings appears to be of significance to research in constructional semantics, yet for another reason. That is, it is likely to shed some light on the direction of the semantic extensions within constructional networks and hence facilitate the task of identifying the different types of links holding between different senses of polysemous forms, another notoriously challenging task in constructional research. Hence, a considerable amount of work is needed before the goal of arriving at a cognitively plausible (and empirically verifiable) theory of post-verbal complement constructions can be regarded as fully accomplished. The present research is one step towards that end.

References

- Achard, Michel. (1998) *Representation of cognitive structures: Syntax and semantics of French sentential complements*. Berlin: De Gruyter Mouton.
- Bolinger, Dwight. (1968) *Aspects of language*. New York: Harcourt, Brace, and World.
- Bresnan, Joan. (1979) *Theory of complementation in English syntax*. New York: Garland.
- Cristofaro, Sonia. (2008) "A constructionist approach to complementation: Evidence from Ancient Greek". *Linguistics* 46(3): 571–606.
- Croft, William. (2001) *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.
- Croft, William and Alan Cruise. (2004) *Cognitive linguistics*. Cambridge: Cambridge University Press.
- de Smet, Hendrik and Hubert Cuyckens. (2005) "Pragmatic strengthening and the meaning of complement constructions: The case of *like* and *love* with the *to*-infinitive". *Journal of English linguistics* 33(1): 3–34.
- Diessel, Holger. (2015) "Usage-based construction grammar". Ewa Dąbrowska and Dagmar Divjak, eds. *Handbook of cognitive linguistics*. Berlin: De Gruyter Mouton, 295–320.
- Dirven, René. (1989) "A cognitive perspective on complementation". Dany Jaspers, Wim Klooster, Yvan Putseys, and Pieter Seuren, eds. *Sentential complementation and the lexicon*. Dordrecht: Foris Publications, 113–39.

- Duffley, Patrick J. (2006) *The English gerund-participle: A comparison with the infinitive*. New York: Peter Lang.
- Egan, Thomas. (2008) *Non-finite complementation: A usage-based study of infinitive and -ing clauses in English*. Amsterdam: Rodopi.
- Geeraerts, Dirk. (2010) "The doctor and the semantician". Dylan Glynn and Kerstin Fischer, eds. *Quantitative methods in cognitive semantics: Corpus-driven approaches*. Berlin: De Gruyter Mouton, 63–76.
- Givón, Talmy. (1980) "The binding hierarchy and the typology of complements". *Studies in language* 4: 333–77.
- . (2001) *Syntax: An introduction*. Vol. 2. Amsterdam: John Benjamins.
- Goldberg, Adele. (1995) *Constructions: A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.
- . (2006) *Constructions at work*. Oxford: Oxford University Press.
- Gries, Stefan Th. and Anatol Stefanowitsch. (2004) "Extending collocation analysis: A corpus-based perspectives on 'alternations'". *International journal of corpus linguistics* 9(1): 97–129.
- Horie, Kaoru. (2000) "Introduction". Kaoru Horie, ed. *Complementation: Cognitive and functional perspectives*. Amsterdam: John Benjamins, 1–10.
- Kaleta, Agnieszka. (2014) *English sentential complementation: A usage-based approach*. Piotrków Trybunalski: Naukowe Wydawnictwo Piotrkowskie.
- . (2021) "How many moods are there in Polish: The case of the Polish subjunctive". *Cognitive semantics* 7(2): 258–89.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. (2014) "The Sketch Engine: Ten years on". *Lexicography* 1: 7–36.
- Langacker, Ronald W. (1987) *Foundations of cognitive grammar*. Vol. 1. *Theoretical prerequisites*. Stanford: Stanford University Press.
- . (1991) *Foundations of cognitive grammar*. Vol. 2. *Descriptive application*. Stanford: Stanford University Press.
- . (1999) "Assessing the cognitive linguistic enterprise". Theo Jansen and Gisela Redeker, eds. *Cognitive linguistics: Foundations, scope, and methodology*. Berlin: De Gruyter Mouton, 23–60.
- . (2010) "Control and the mind/body duality: Knowing vs. effecting". Elżbieta Tabakowska, Michał Choński, and Łukasz Wiraszka, eds. *Cognitive linguistics in action*. Berlin: De Gruyter Mouton, 165–208.
- Lemmens, Martin. (2015) "Cognitive semantics". Nick Riemer, ed. *Routledge handbook of semantics*. London: Routledge, 90–105.
- Michaelis, Laura A. (2004) "Type shifting in construction grammar: An integrated approach to aspectual coercion". *Cognitive linguistics* 15(1): 1–67.
- Michaelis, Laura A. and Josef Ruppenhofer. (2001) *Beyond alternations: A construction-based approach to the applicative pattern in German*. Stanford: CSLI Publications.

- Noonan, Michael. (1985) "Complementation". Timothy Shopen, ed. *Language typology and syntactic description*. Vol. 2. *Complex constructions*. Cambridge: Cambridge University Press, 42–140.
- Ransom, Evelyn R. (1986) *Complementation: Its meanings and forms*. Amsterdam: John Benjamins.
- Rudanko, Juhani. (2017) *Infinitives and gerunds in recent English: Studies of non-finite complements with data from large corpora*. Basingstoke: Palgrave Macmillan.
- Ruohonen, Juho and Juhani Rudanko. (2020) *Infinitival vs. gerundial complementation with afraid, accustomed, and prone: Multivariate corpus studies*. Basingstoke: Palgrave Macmillan.
- Searle, John. (1979) *Expression and meaning: Studies in the theory of speech acts*. Cambridge: Cambridge University Press.
- Smith, Michael B. (2008) "The semantics of complementation in English: A cognitive semantic account of two English complement constructions". *Language sciences* 31: 360–88.
- Stefanowitsch, Anatol and Stefan Th. Gries. (2003) "Collostructions: Investigating the interaction between words and constructions". *International journal of corpus linguistics* 8(2): 209–43.
- Talmy, Leonard. (1988) "Force dynamics in language and cognition". *Cognitive science* 12: 49–100.
- Taylor, John. (2019) "Prototype effects in grammar". Ewa Dąbrowska and Dagmar Divjak, eds. *Cognitive linguistics—Key topics*. Berlin: De Gruyter Mouton, 127–47.
- Tomasello, Michael. (2003) *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
- Wierzbicka, Anna. (1988) *The semantics of grammar*. Amsterdam: John Benjamins.
- Wolff, Phillip. (2003) "Direct causation in the linguistic coding and individuation of causal events". *Cognition* 88(1): 1–48.
- Yoon, Jiyoung and Stefanie Wulff. (2016) "A corpus-based study of infinitival and sentential complement constructions in Spanish". Jiyoung Yoon and Stefan Th. Gries, eds. *Corpus-based approaches to Construction Grammar*. Amsterdam: John Benjamins, 145–64.

Agnieszka Kaleta
Piotrków Academy
Piotrków Trybunalski, Poland
agnieszka.kaleta@apt.edu.pl

Binding in South Slavic and DP: A Data-Driven Approach

Ivana LaTerza, Petya Osenova, and Boban Karapejovski

Abstract: This paper reports on a set of experiments designed to test the binding potential of prenominal possessives in Bulgarian, Macedonian, and Serbian. Despić (2013) argues that the differences in binding possibilities observed between English and Serbian provide support for the Parameterized DP Hypothesis (e.g., Fukui 1988; Zlatić 1997; Bošković 2003, 2005, 2008). LaTerza (2016) tests whether the claim holds true for two South Slavic DP-languages, Bulgarian and Macedonian, and concludes that it does not. Data provided in LaTerza 2016 is further discussed in Franks 2019. Based on three interesting observations—the use of a clitic vs. full pronoun, different binding behavior of pronominal and nominal possessives in Bulgarian, and acceptability judgments reported for Macedonian and Serbian—Franks (2019) concludes that Bulgarian and Macedonian have the same binding potentials as English, confirming Despić’s original hypothesis. Srdanović and Rinke (2020) provide Serbian experimental data focusing on possessives in subject position and coreferential readings of pronouns in object positions. The authors show that Serbian allows coreferential readings just like English, especially when clitics are used. Our paper provides experimental data for Bulgarian, Macedonian, and Serbian. Our conclusions are that the three languages exhibit almost identical binding potentials. This finding is in line with the ones in Srdanović and Rinke 2020 since it also disproves the claim that the differences in binding result from the nominal structure present in a language: DP or NP.

1. Debate on DP and Binding

The universality of a Determiner Phrase (DP) as a functional projection in the nominal domain has been questioned ever since it was introduced (Abney 1987¹). Fukui (1986) proposed that the DP was not universal based on his observation that Japanese lacks articles. A decade later, Zlatić (1997) revived and further developed Fukui’s idea, arguing that the sole presence of articles in a language indicates the presence of a DP. These proposals initiated the discussion on parametric variation of DP. Two opposing views emerged: (i) the Universal DP Hypothesis (UDPH), which holds that DP projects in all

¹ Abney’s work is based on the previous work of Brame (1982), Szabolcsi (1983), and Fukui and Speas (1986).

languages (Progovac 1998; Rappaport 2001; Bašić 2004; Cinque 2005, among others), and (ii) the Parameterized DP Hypothesis, which claims that DP projects only in languages with (definite) articles (Fukui 1986; Corver 1992; Zlatić 1997; Bošković 2005; Despić 2011).

The Parameterized DP Hypothesis has been most fully developed and explored in the works of Bošković and his followers, with a focus on Serbian (Stjepanović 1998; Bošković 2003, 2005; Trenkić 2004; Despić 2011; Talić 2013). This view specifically holds that the presence of DP correlates exclusively with the presence of a definite article.² Since Serbian lacks definite articles, it is claimed to lack DP. With no DP available, D-like elements are argued to be NP-adjoined. Such nominal structure has been claimed to have numerous empirically verified syntactic implications³ (Bošković 2008, 2012; Despić 2013). This paper investigates one such implication: binding.

Despić (2009, 2011, 2013)⁴ observes that English and Serbian prenominal possessives differ in binding possibilities: English allows coreferential readings with R-expressions or pronouns elsewhere in the clause ((1a) and (1b), respectively), whereas Serbian does not, (2).

- (1) a. His_i father considers John_i highly intelligent.
- b. John_i's father considers him_i highly intelligent.

(Despić 2009: 20, (3–4))

² The privileged status given to one specific element—the definite article—raises a number of questions: (i) language acquisition of a D category; (ii) language variation (North Frisian and Faroese have multiple lexical items corresponding to the English definite article) (Delsing 1993; Julien 2003; Schwarz 2009); (iii) definite articles do not exhibit a unique behavior among other determiners in English (Jackendoff 1977; Chomsky 1981); (iv) the proposed division among determiners has no semantic anchoring (Barwise and Cooper 1981; Heim 2002); and (v) the absence of the definite article entails that all the structure that the item might be responsible for licensing, such as relative clauses, should be missing as well (Smith 1964; Vergnaud 1974). See LaTerza 2014 for discussion.

³ Thorough cross- and intra-linguistic investigation has, however, shown that some of these implications need to be re-examined (Rappaport 2001; Bašić 2004; Runić 2006; Ivšić 2008; Caruso 2011; Bailyn 2012; Schoorlemmer 2012; Pereltsvaig 2013; Stanković 2013; Dubinsky and Tasseva-Kurktchieva 2014, etc.).

⁴ Note that while Despić uses the same methodology for observing Serbian and English data, the experiments reported in this paper lack the baseline experimental data for English that would be directly comparable with our data for Bulgarian, Macedonian, and Serbian. Thus we take it for granted that the cited sentences for English in Despić 2009 are grammatical also in our context. We would like to thank one of our anonymous reviewers for pointing out this very important issue.

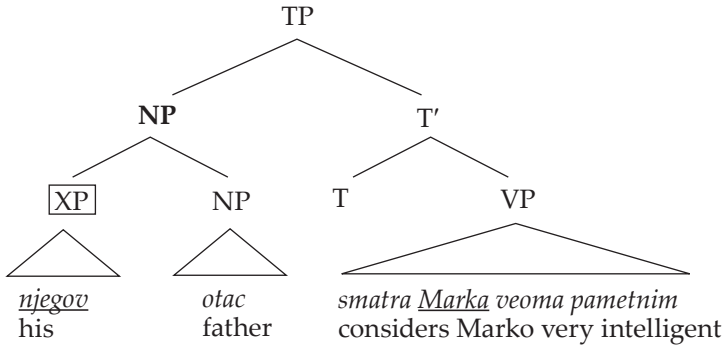
- (2) a. *Njegov_i otac smatra Marka_i veoma pametnim.
his father considers Marko very intelligent
'His_i father considers Marko_i very intelligent.'
- b. Markov_i otac smatra njega_i veoma pametnim.
Marko.POSS father considers him very intelligent
'Marko_i's father considers him_i very intelligent.'
- (Serbian; Despić 2009: 22, (11–12))

(3)

```
graph TD
    TP --> DP
    TP --> T_prime[T']
    DP --> D[D]
    DP --> PossP1[PossP]
    D --> his[his]
    PossP1 --> NP1[NP]
    NP1 --> his2[his]
    PossP1 --> PossP2[PossP]
    PossP2 --> s['s]
    PossP2 --> NP2[NP]
    NP2 --> father[father]
    T_prime --> T[T]
    T --> triangle1[triangle]
    T_prime --> VP[VP]
    VP --> considers[considers]
    VP --> complementizer[John very intelligent]
```

⁵ Note also that Despić (2011: 133, (54)) argues that in English full possessors are in SpecDP, while pronominal possessors are in SpecPossP, with a caveat, listed in a footnote, that the full DP possessor might move from SpecPossP to SpecDP. Such a structure would predict that full possessor DPs would be able to bind outside of DP in English, contrary to judgments reported in (1b).

(4)



(LaTerza 2016: 743, (5))

Such a proposal predicts that two Slavic languages with definite articles, Bulgarian and Macedonian, should parallel English with respect to binding. LaTerza (2016) tests this prediction and concludes that it does not hold true: both Bulgarian and Macedonian, on par with Serbian, disallow relevant coreferential readings.^{6 7 8}

⁶ LaTerza (2016: 748, fn. 13) also reports that one Bulgarian speaker has different acceptability judgments regarding example (5) and points out that the observed variation calls for a controlled study.

⁷ Note that the Bulgarian and Macedonian examples differ somewhat from the Serbian examples taken from Despić 2009. In particular, they do not contain the verb *smatrati* 'consider', which might involve a small clause structure and therefore interfere with Condition B. LaTerza (2016: 742, fn. 4) addresses this fact and further notes that Despić (2011, 2013) provides examples with other verbs with no change in acceptability. She provides only examples *without* a small clause structure in order to avoid any interference it might have with the relevant binding potentials.

⁸ Despić (2015) argues that languages with postnominal articles share certain characteristics with languages with no articles rather than languages with pre-nominal articles, such as the presence of reflexive pronouns. Talić (2020) makes a similar observation with respect to the presence of bare lexical projections.

To account for the presence of reflexive pronouns in Bulgarian but not English, Despić (2011) assumes that possessors move to the edge of D because D has some sort of Edge feature. These assumptions apply to all languages with postnominal definiteness marking (Icelandic, Faroese, Swedish, Danish, Norwegian, Bulgarian, Macedonian, Romanian). Note that such a structure would allow for possessors in these languages to bind out of DP. Our findings for Bulgarian and Macedonian are consistent with this observation. Note, however, that as far as binding potentials are concerned, the division between DP- and NP-languages would then not be based solely on the nominal structure (DP or no DP) but also on some Edge Feature on D that would trigger the movement of possessors to SpecDP. Further investigation of binding possibilities in other languages with postnominal definiteness marking would be highly

- (5) a. *Negovijat_i papagal uxapa Ivan_i včera. (Bulgarian)
his.DEF parrot bit Ivan yesterday
Intended: 'His_i parrot bit Ivan_i yesterday.'
- b. *Ivanovijat_i papagal nego_i uxapa včera.
Ivan.POSS.DEF parrot him bit yesterday
Intended: 'Ivan_i's parrot bit him_i yesterday.'
(LaTerza 2016: 748, (13))
- (6) a. *Negoviot_i papagal go_i grizna Jovan_i (Macedonian⁹)
his.DEF parrot him.CL bit Jovan
včera.
yesterday
Intended: 'His_i parrot bit Jovan_i yesterday.'
- b. *Jovanoviot_i papagal go_i grizna nego_i včera.
Jovan.POSS.DEF parrot him.CL bit him yesterday
Intended: 'Jovan_i's parrot bit him_i yesterday.'
(LaTerza 2016: 748, (14))

Franks (2019) brought up three interesting points with respect to the binding data, as presented in LaTerza 2016: (i) the use of clitic and full pronoun forms in Bulgarian and how they interact with binding possibilities, (ii) the difference in acceptability of pronominal and nominal possessives in Bulgarian, and (iii) the overall acceptability of relevant coreferential readings in Macedonian and Serbian. We will address each of these in turn below.

2. Clitic vs. Full Pronoun

Franks (2019) has made the interesting empirical claim that the relevant binding data turn crucially on whether clitic or full pronoun forms are employed, a point which, if correct, would have serious implications for how binding arguments are evaluated. In particular, Franks reports that LaTerza's Bulgarian example containing a full pronoun, (7a), allows relevant coreferential reading

informative in this regard. Note that Franks's (2019) interpretation of Despić 2011 is challenged by our findings.

⁹ Macedonian examples, unlike Bulgarian, contain an object clitic, which is required when the object is definite. Please note that clitic doubling exists in Bulgarian as well and it mainly depends on the information structure. Such structures have not been examined here nor in the works cited. See Nicolova 1998 (esp. pp. 151–55) for structured information on clitic doubling in Bulgarian.

if a clitic is used instead, as in (7b). Hence, he concludes, LaTerza's argument that Bulgarian and English binding potentials differ is undermined.

- (7) a. *Ivanovijati_i papagal nego_i uxapa včera. (Bulgarian)
 Ivan.POSS.DEF parrot him bit yesterday

Intended: 'Ivan_i's parrot bit him_i yesterday.'

(LaTerza 2016: 748, (13))

- b. Ivanovijati_i papagal go_i uxapa včera.
 Ivan.POSS.DEF parrot him bit yesterday

'Ivan_i's parrot bit him_i yesterday.'

(Franks 2019: 70, (18))¹⁰

Interestingly, Serbian, like Bulgarian, has a possibility of using either a full pronoun or clitic. And, in fact, the pronoun used in Despić's example, (2b), is a full pronoun. LaTerza (2016) kept the same structure that Despić (2009) used in his original paper where he made the claims about binding possibilities in Serbian vs. English. In that respect then, Franks (2019) is arguing against LaTerza's data without subjecting the original Despić's data to the same standard.

Despić (2009: 22, fn. 4) addresses the question of full pronoun vs. clitic used in his examples, and claims that the use of full vs. clitic form of a pronoun does not interfere with the acceptability of the relevant coreference in Serbian. He further explains that when a clitic is used, "the sentence somehow 'improves' (but still stays ungrammatical)" (ibid.). Despić ascribes the observed improvement to the cross-linguistic observation that full pronouns generally introduce new referents. Therefore, a full pronoun, e.g., *njega* 'him', cannot refer to an already introduced referent, e.g., *Markov* 'Marko's'. He further claims that this observation, in addition to the violation of Condition B, makes the relevant example "more ungrammatical" when a full pronoun is used instead of a clitic. So, in Serbian, the type of pronoun used does not seem to interfere with the binding potentials. Therefore, if the binding differences between the two types of languages stem from the presence or absence of DP, the puzzling difference in the use of full and clitic pronouns in Bulgarian remains to be accounted for.

Srdanović and Rinke (2020) present Serbian experimental data addressing the issue of full vs. clitic form in relevant structures and conclude that clitics are more likely to be interpreted as coreferential (55%) than full pronouns (41%). The differences in binding potentials cannot be attributed to the nominal structures of the languages, DP or NP, but rather the availability of different types of pronouns in a language. The authors favor the pragmatic-based approach, i.e., discourse conditions, to coreference potential over the

¹⁰ Word-for-word glosses have been modified to adhere to the JSL stylesheet.

syntactic one. The results of this study are in opposition to the findings reported in LaTerza 2016, but what unifies the two is the fact that both show that the nominal structure cannot be responsible for the binding potentials in the languages in question.

It is important to note that the issue of full vs. clitic pronouns as discussed in Franks 2019 does not seem to undermine the LaTerza squib nearly as much as it claimed, since the full pronoun data, (7a), is not in dispute and still argues against Despić. If binding potentials directly relate to the presence or absence of DP in the nominal structure, where DP allows for coreferential readings, the Bulgarian example in (7a) remains to be explained under this theory.

3. Pronominal vs. Nominal Possessives

Another observation made in Franks 2019 concerns the difference in acceptability of pronominal and nominal possessive coreferential readings in Bulgarian. LaTerza (2016) reports that Bulgarian coreferential readings between a *pronominal* possessive and R-expression, (8a), and a *nominal* possessive and pronoun, (8b), are equally unacceptable.

- (8) a. *Negovijati_i papagal uxapa Ivan_i včera. (Bulgarian)
 his.DEF parrot bit Ivan yesterday
 Intended: 'His_i parrot bit Ivan_i yesterday.'
- b. *Ivanovijati_i papagal nego_i uxapa včera.
 Ivan.POSS.DEF parrot him bit yesterday
 Intended: 'Ivan_i's parrot bit him_i yesterday.'
- (LaTerza 2016: 748, (13))

As discussed in §2 above, Franks (2019) reports that (8b) is unacceptable because of the full pronoun, and as soon as the clitic is used instead, the example becomes completely acceptable for the relevant coreferential readings. Example (8a), on the other hand, presents a puzzle for him since "some B[ul]g[arian] speakers, e.g., Iliyana Krapova (p.c.), do concur with the judgment LaTerza reports" (Franks 2019: 73).

Franks continues to show that by introducing different types of demonstratives and quantifiers in the structure, such as *tezi* 'these' in (9), the coreferential readings become available.¹¹

¹¹ A word of caution regarding examples used to illustrate this point: the example with the demonstrative, (9), is in plural, whereas the example without the demonstrative, (8a), is in singular. It is always good practice to keep the examples minimally different since other factors might be interfering with relevant interpretations.

- (9) Tezi negovi_i papagali uxapaxa Ivan_i včera. (Bulgarian)
 these his parrots bit Ivan yesterday
 ‘These parrots of his_i bit Ivan_i yesterday.’ (Franks 2019: 74, (23))

He then concludes that “[i]t is unclear why having just a DP above NP, as in [(8a)], instead of a DP and a QP, is not sufficient to override the R-expression effect (since replacing the R-expression with a clitic pronoun makes these good for Condition B)... The solution surely has to do with the depth of nominal structure, although just how to calculate that depth remains a puzzle. Nonetheless, regardless of how such subtleties are explained, these data demonstrate that demonstratives and quantifiers indeed count as adding a distinct category above NP” (Franks 2019: 75). According to this theory, it follows that demonstratives are **not** in DP in Bulgarian but rather in QP, an assumption that needs to be addressed since demonstratives are natural candidates for DP projection across languages.

The Bulgarian data that Franks discusses actually clearly demonstrate that DP by itself cannot be the explanation of binding potentials, a point which he takes to be crucial for DP- and NP-language binding potentials. In other words, his work shows that something other than the presence or absence of DP drives the relevant coreferential readings, at least in Bulgarian.

4. Acceptability of Relevant Coreferential Readings

Another point discussed in Franks 2019 is the overall acceptability of relevant coreferential readings in Macedonian and Serbian. In particular, Franks reports that all Macedonian speakers he consulted find coreferential readings in the examples reported in LaTerza 2016 viable (cf. (6)):

Furthermore, other examples Franks uses to support this point contain a ‘make’-causative, the structure that under a variety of analyses (Pesetsky 1994; Cheung and Larson 2018) is claimed to involve *derived* subjects. One of the key properties of *derived*-subject constructions is that they interact with binding relations, permitting backward binding of anaphors (cf. [*Her_i*; many problems] made *Mary_i* uneasy). In other words, it is exactly the kind of structure one does not want to use in this context since other factors might be interfering with binding potentials.

- (i) a. *Nejnite_i problemi pritesnjava Maria_i mnogo.
 her.DEF problems troubled Maria much
 Intended: ‘Her_i problems made Maria_i very uneasy.’
 b. Tezi nejni_i problemi pritesnjava Maria_i mnogo.
 these her problems troubled Maria much
 ‘These problems of hers_i made Maria_i very uneasy.’ (Franks 2019: 74, (25))

- (10) a. Negoviot_{i/j} papagal go_i grizna Jovan_i (Macedonian)
 his.DEF parrot him bit Jovan
 včera.
 yesterday
 ‘His_{i/j} parrot bit Jovan_i yesterday.’
- b. Jovanoviot_i papagal go_{i/j} grizna (nego_{i/j}) včera.
 Jovan.POSS.DEF parrot him bit him yesterday
 ‘Jovan_i’s parrot bit him_{i/j} yesterday.’ (Franks 2019: 72, (14’))

Franks (2019: 71) explains that the initial interpretations his consultants report align with the reports made in LaTerza 2016, but that “all speakers readily concede that [the owner of the parrot and the person bitten] could be the same”. He states, “[i]t is thus possible that LaTerza was collecting preferred/dominant readings rather than absolute judgments, which could explain the discrepancy between what she reports and the judgments I obtained. The fact nonetheless remains that the coreference possibilities in Mac[edonian] are exactly as they are in English, which is precisely what we expect if Mac[edonian], like English, is a DP-language” (Franks 2019: 72).

Franks draws a distinction between “preferred/dominant reading and absolute judgment” without explaining what the absolute judgment means. Binding data in particular seems susceptible to variation, so it remains unclear what the absolute judgment would mean in this context.¹²

Franks (2019: 63, fn. 4) also discusses the acceptability of such coreferential readings in English and says “that English speakers do not hesitate in accepting the intended translations.” This statement is in opposition to the one made about Macedonian consultants since their initial interpretations are the ones that do not involve coreference.

Unlike Macedonian, Franks (2019: 70, (18)) reports that Serbian binding data is exclusive, that is, the “speakers consistently disallow coreference in comparable sentences”.¹³

¹² Franks mentions that the possessive adjective derived from the R-expression *Jovan* in Macedonian, *Jovanoviot*, is not particularly natural, and that the more natural way to express this is to use the prepositional phrase *na Jovan* ‘of Jovan’. When the prepositional variant is used, he adds, “no Mac[edonian] speakers expressed any reservations about the acceptability of coreference” (Franks 2019: 72, fn. 8). This is not surprising given that the R-expression is inside a prepositional phrase, and as such, it does not c-command the relevant pronoun, rendering coreference possible.

¹³ Example (11) is a Serbian equivalent of Franks’s Bulgarian example (18) (Franks 2019: 70).

- (11) *Jovanov_i papagaj ga_i je juče ugrizao. (Serbian)
 Jovan.POSS parrot him AUX yesterday bit
 Intended: 'Jovan_i's parrot bit him_i yesterday.'

It could be the case that the absolute judgment refers to consistent disallowing or allowing of coreferential readings. One very important point to keep in mind here is the fact that the data reported in LaTerza 2016, and then further discussed in Franks 2019, stem from a handful of speakers, many of whom have linguistic training and are familiar with the theories tested when asked for native speaker judgments. To settle this issue and provide unbiased data, we ran a controlled study of relevant binding data in Bulgarian, Macedonian, and Serbian. The design and results of the study are presented in the next section.

5. Experimental Study

Our study tested how native speakers of Bulgarian, Macedonian, and Serbian interpret pronouns and pronominal possessives in the following scenarios: (i) nominal possessives in subject positions and pronouns in object positions, (12a), and (ii) pronominal possessives in subject positions and R-expressions in object positions, (12b). In other words, we wanted to see if the speakers readily choose coreferential interpretations available within a clause—i.e., coreferential readings with R-expressions or pronouns, (12a) and (12b), respectively—or outside a clause (some referent outside of the clause).

- (12) a. Ivanov_i papagaj ga_i je ugrizao. (Serbian)
 Ivan.POSS parrot him AUX bit
 'Ivan_i's parrot bit him_i.'
- b. Njegov_i papagaj je ugrizao Ivana_i.
 his parrot AUX bit Ivan
 'His_i parrot bit Ivan_i.'

All items (critical and fillers) in all three languages (Bulgarian, Macedonian, and Serbian) had the exact same format and conditions.¹⁴

¹⁴ To our knowledge, there is no other study reported in the literature testing these structures in Bulgarian, Macedonian, and Serbian.

5.1. Design

We devised sets of sentences, each testing one of the relevant interpretations: pronouns and pronominal possessives. The scenarios presented included a sentence with relevant binding elements, followed by a question on interpretation of the relevant element: pronoun or pronominal possessive, (13a) and (13b), respectively:

(13) a. Pronoun interpretation

Lukin papagaj ga je ugrizao. (Serbian)
 Luka.POSS parrot him AUX bit
 Koga je Lukin papagaj ugrizao?
 whom AUX Luka.POSS parrot bit
 ‘Luka’s parrot bit him. Who did Luka’s parrot bite?’

b. Pronominal possessive interpretation

Njegov papagaj je ugrizao Luku.
 his parrot AUX bit Luka
 Čiji papagaj je ugrizao Luku?
 whose parrot AUX bit Luka
 ‘His parrot bit Luka. Whose parrot bit Luka?’

The participants were given two possible answers to choose from: (i) the referent mentioned in the relevant context (Luka) and (ii) someone else (not the referent).

We also included corresponding examples with two overtly mentioned referents to see if the overt presence of another referent in the relevant context changes the preferred interpretation. The second referent was always outside of the critical clause.

(14) a. Pronoun interpretation – two overt referents (Serbian)

Luka ima papagaja. Marko se igra s njim.
 Luka has parrot Marko REFL plays with it
 Lukin papagaj ga je ugrizao.
 Luka.POSS parrot him AUX bit
 Koga je Lukin papagaj ugrizao?
 whom AUX Luka.POSS parrot bit
 ‘Luka has a parrot. Marko is playing with it. Luka’s parrot bit him. Who did Luka’s parrot bite?’

(14) b. Pronominal possessive interpretation – two overt referents

Luka ima papagaja. Marko se igra s njim.
 Luka have parrot Marko REFL play with it

Njegov papagaj je ugrizao Luku.
 his parrot AUX bit Luka

Čiji papagaj je ugrizao Luku?
 whose parrot AUX bit Luka

‘Luka has a parrot. Marko is playing with it. His parrot bit Luka.
 Whose parrot bit Luka?’

In these cases, the participants were given three options to choose from: referent 1 (Luka), referent 2 (Marko), or someone else.¹⁵

There were six examples for each of the four conditions: (i) one overt referent, pronoun interpretation; (ii) one overt referent, pronominal possessive interpretation; (iii) two overt referents, pronoun interpretation; and (iv) two overt referents, pronominal possessive interpretation. There were sets of minimally different sentences for all four conditions (as shown in (13) and (14) above). All three languages had the same sets to avoid any potential semantic differences among them. See the appendix for the list of all critical items for all three languages.

The examples testing pronominal interpretations in all three languages included clitics and **not** full pronouns, in order to test the claims made in Franks 2019, as discussed in §2 above. Our study complements and further builds on Srdanović and Rinke 2020 by testing coreferential interpretations of clitics in Serbian¹⁶ and adding Bulgarian and Macedonian data.

¹⁵ A reviewer points out that participants should have been given another option, that is, “coreferent OR someone else”. We ran such an experiment with the same data and found that participants in all three languages either choose the OR option or the responses are almost evenly spread among the provided possible answers. We wanted to get at preferences, so we ran a follow-up experiment excluding the OR option, and the results of that study are reported in this paper. There are other possible follow-up experiments that could provide more insight into this topic, such as giving participants a scale rather than an either/or choice. There could be two ways to do this: either (i) for a sentence like *Luka’s parrot bit him*, the scale goes from “*him* = Luka” to “*him* = someone else”; or (ii) asking for acceptability judgements, e.g., on a scale from “very acceptable” to “unacceptable”, participants would judge items like *Luka’s parrot bit him*, where *him* = Luka, and *Luka’s parrot bit him*, where *him* = someone else.

¹⁶ Srdanović and Rinke (2020) used a picture selection task where participants were presented with a context and a critical sentence (either with a full pronoun or a clitic) and then were shown two pictures corresponding to two different coreferential readings and asked to choose the correct one.

All the verbs used in critical items were common transitive verbs to avoid any other potential structural interference with respect to binding interpretations. We used frequent lexical items and avoided using distractors, such as nominal or verbal modifiers. Each participant was presented with all conditions, i.e., all 24 critical items. We also included 24 fillers, which followed the same pattern of having to choose one of the provided answers. All the items were randomized. The platform used to distribute the questions and collect responses in all three languages was Qualtrics.

5.2. Participants

There were 12 participants in the Bulgarian study, 17 in Macedonian, and 16 in Serbian. All participants were adult native speakers with no prior linguistic training. The participants were not paid. The study was anonymous. The recruitment of the participants was done via social media and distribution of the survey link to friends and acquaintances. At the beginning of the study, all participants were asked to verify that they are 18 years or older and that they are native speakers of the language in question.

5.3. Results

The results reveal that overall participants have preference for non-binder interpretation in all three languages in all conditions. In particular, non-binder interpretations were selected 80.9% of the time in Bulgarian, 66.66% in Macedonian, and 65.1% in Serbian.

However, in all three languages, the preferences differed as a function of pronominal possessive adjective and pronoun, such that non-binder interpretations were chosen more often for pronouns than for pronominal possessive adjectives. The results for each language are shown in Table 1 below.

Table 1. Percentages of chosen non-binder interpretations for pronominal possessive adjectives and pronouns

Non-binder interpretations	Bulgarian	Macedonian	Serbian
Pronominal possessive adjective	75.69	66.17	57.29
Pronoun	86.1	67.15	72.91

The difference between the pronominal possessive adjective and pronoun non-binder interpretation in Bulgarian is 10.41%, Macedonian 0.98%, and

Serbian 15.62%. Macedonian shows the least difference between the two conditions, suggesting that the preference for non-binder interpretation for pronominal possessive adjectives and pronouns seems to be minimally different. Bulgarian and Serbian, on the other hand, show bigger differences.

Similarly, in all three languages, the preferences differed as a function of having one overtly introduced referent and two overtly introduced referents, such that non-binder interpretations were chosen more often for one overtly introduced referent. The results for each language are shown in Table 2 below.

Table 2. Percentages of chosen non-binder interpretations in the context of one overtly introduced referent and two overtly introduced referents for pronominal possessive adjectives and pronouns

Non-binder interpretations	Bulgarian	Macedonian	Serbian
One overtly introduced referent	87.5	73.52	72.39
Two overtly introduced referents	74.3	59.8	57.81

The difference between non-binder interpretations in the context of one overtly introduced referent and two overtly introduced referents is 13.2% in Bulgarian, 13.72% in Macedonian, and 14.58% in Serbian. The differences observed for all three languages are minimally different, suggesting that the three languages behave very similarly in this respect: speakers prefer non-binding interpretations more often when there is one overtly introduced referent than when there are two.

Next, in Bulgarian and Macedonian, the preferences differed as a function of having one overtly introduced referent and two overtly introduced referents with pronominal possessive adjective, such that non-binder interpretations were chosen more often for pronominal possessive adjectives in the context of one overtly introduced referent. Serbian shows no difference between these two conditions. The results for each language are shown in Table 3.

Table 3. Percentages of chosen non-binder interpretations for pronominal possessive adjectives in the context of one overtly introduced referent and two overtly introduced referents

Non-binder interpretations	Bulgarian	Macedonian	Serbian
One overtly introduced referent	77.77	70.58	57.29
Two overtly introduced referents	73.61	61.76	57.29

The difference between non-binder interpretations for pronominal possessive adjectives in the context of one overtly introduced referent and two overtly introduced referents is 4.16% in Bulgarian and 8.82% in Macedonian. Serbian shows no difference, suggesting that speakers uniformly prefer non-binder interpretation for pronominal possessive adjectives, regardless of the number of overtly present referents—one or two.

Lastly, in all three languages, the preferences differed as a function of having one overtly introduced referent and two overtly introduced referents with pronouns, such that non-binder interpretations were chosen more often for pronouns in the context of one overtly introduced referent. The results for each language are shown in Table 4 below.

Table 4. Percentages of chosen non-binder interpretations for pronouns in the context of one overtly introduced referent and two overtly introduced referents

Non-binder interpretations	Bulgarian	Macedonian	Serbian
One overtly introduced referent	97.2	76.47	87.5
Two overtly introduced referents	75	57.84	58.3

The difference between non-binder interpretations for pronouns in the context of one overtly introduced referent and two overtly introduced referents in Bulgarian is 22.2, in Macedonian 18.63, and in Serbian 29.2. This result aligns with the observed preference for non-binder interpretations for pronouns on the one hand and the context of one overtly introduced referent on the other.

To sum up, the results show an overall preference for non-binder interpretation in all conditions in all three languages. However, the preference differed as a function of (i) pronoun and pronominal possessive adjective and (ii) having one overtly introduced referent and two overtly introduced referents, such that non-binder interpretations were chosen more often for the former rather than the latter (pronouns, one overtly introduced referent). When the two conditions intersected, the results show preference for non-binder interpretation in the context of one overtly introduced referent for pronouns in all three languages.

6. Discussion

The results of our study show that there is an overall preference for non-binder interpretations in all conditions we tested in all three languages. This finding

invalidates the argument that the differences in binding possibilities originally observed for English and Serbian stem from the nominal structure of the language: the presence of DP makes it impossible for pronominal possessives to c-command outside DP, allowing for coreferential readings, whereas the lack of DP allows them to c-command outside of their NP since they are NP-adjoined. If the binding potentials are determined by the presence or absence of DP, then we would expect to see differences between Serbian (an NP-language), on the one hand, and Bulgarian and Macedonian (DP-languages), on the other. We have not found evidence for this in our study.

There was an overall higher percentage of non-binder interpretations chosen for pronouns versus pronominal possessive adjectives. This finding could be attributed to the well-known empirical observation that backward anaphora between a pronoun and a following R-expression is blocked when the R-expression bears focus (Chomsky 1976; Williams 1997; Erteschik-Shir 1997; Bianchi 2010):

- (15) a. *His_i wife loves JOHN_i.
 b. His_i wife LOVES John_i. (Bianchi 2010: 9, (6), (7))

Reinhart (1986) proposes a topic-antecedent hypothesis to account for the contrast shown in (15):

- (16) Backward anaphora is possible only if the antecedent is in sentence-topic position. (Reinhart 1986: 138–40)

The R-expression in (15a) cannot be the antecedent of the backward anaphora, because it bears a new information focus and as such cannot be the sentence topic (Zubizarreta 1998), hence it fails to satisfy (16).

Since our study was conducted online, where participants read the sentences themselves, we could assume that the participants were putting focus on R-expressions and hence opting for non-binder interpretations.

We also observed that non-binder interpretations were chosen more often in the context of one overtly mentioned referent versus two,¹⁷ which could be explained by contrast between the potential referents. In particular, when there is only one overtly mentioned referent, participants are more likely to opt for the interpretation of the relevant elements where their binders are not

¹⁷ We are aware that the results for Serbian in Table 3 show no difference with respect to having one or two referents when possessive adjectives have been used. We will not speculate whether this ignorance to the number of introduced referents in context is a strong tendency in Serbian or just a specific output from the current pool of participants. Obviously, replications of the present survey have to be envisaged as future work.

present in the text. However, when there are two overtly mentioned referents, the context seems to be interpreted as contrastive—two referents are “competing” for the role of a binder. The one that is closer to the element needed to be bound (within the same clause) seems to be favored over the one that is further away.

Srdanović and Rinke’s (2020) study reports on what we refer to as two overtly mentioned antecedents in a pronoun condition for Serbian. Their findings show that non-binding interpretations are chosen 45% of the time, whereas our study shows 58.3%. Even though the findings do not completely align, we can see that in both studies participants allow for both readings but have a slightly different preference for one reading over the other. This difference could be attributed to the techniques used in the two studies: picture-matching (their study) vs. multiple-choice questions (our study), or reading sentences out loud (their study) vs. not being required to read the sentences out loud (our study). But maybe the most important factor is the introductory context. In particular, in Srdanović and Rinke 2020 (p. 172) the potential antecedents are introduced by an existential construction with the relevant possessive: ‘Here are Petar, Petar’s bull, and Jovan.’ The introductory context in our study focuses more on the situation that the participants are in, such as ‘Luka has a parrot. Marko is playing with it.’

Our study also provided invaluable insights into the issues raised by Franks (2019), as discussed in sections 2, 3, and 4 above. First is the issue of full vs. clitic pronominal forms in Bulgarian. The claim is that the coreferential reading in Bulgarian improves when a full pronoun is replaced with a clitic. We included examples that are almost identical to the examples provided in Franks 2019 to test this claim. One such example is the following (cf. (7b)):

- (17) Ivanovijati_i papagal go_i uxapa včera. (Bulgarian)
 Ivan.POSS.DEF parrot him bit yesterday
 ‘Ivan_i’s parrot bit him_i yesterday.’ (Franks 2019: 70, (18))

The only difference between this example cited in Franks 2019 and the example we used in our study, provided in (18) below, is the omission of the adverb *včera* ‘yesterday’. We included this example in both conditions: one overtly mentioned referent, (18a), and two overtly mentioned referents, (18b).

- (18) a. Ivanovijati_i papagal go_i uxapa. (Bulgarian)
 Ivan.POSS.DEF parrot him bit
 Kogo uxapa Ivanovijat papagal?
 who bit Ivan.POSS.DEF parrot
 ‘Ivan_i’s parrot bit him_i. Who did Ivan’s parrot bite?’

- (18) b. Ivan ima papagal. Martin si igrae s nego.
 Ivan has parrot Martin REFL plays with him
 Ivanovijat_i papagal go_i uxapa.
 Ivan.POSS.DEF parrot him bit
 Kogo uxapa Ivanovijat papagal?
 who bit Ivan.POSS.DEF parrot
 ‘Ivan has a parrot. Martin is playing with it. Ivan_i’s parrot bit him_i. Who did Ivan’s parrot bite?’

In both conditions, non-binding interpretations for this particular example were prevalent: for one overtly present referent, non-binder interpretation was chosen 97.2% of the time, and for two overtly present referents, non-binder interpretation was chosen 75% of the time. In addition to this example, we used multiple others to test the claim and observed that non-binder interpretations are favored. This suggests that Bulgarian binding potentials do not seem to relate to the presence of clitic vs. full pronouns, where the clitic makes binding interpretations more viable. Bulgarian, therefore, does not align with English, disproving the claim made in Franks 2019.

This finding ties into the second issue raised in Franks 2019 regarding the difference in acceptability of pronominal and nominal possessive coreferential readings in Bulgarian. Since coreferential readings between a nominal possessive and pronoun have a prevalent non-binding interpretation, there does not seem to be any difference in acceptability between the two structures, contra Franks 2019. Whether it is a structure where coreferential readings are tested between (i) a nominal possessive and pronoun, as in (18), or (ii) a pronominal possessive and R-expression, as shown in (8a), the study shows that non-binder interpretations are preferred in both, with minimal difference between them: for (i) 86.1% and for (ii) 75.69%.

Finally, our study shows that binding data interpretations vary among speakers of all three languages tested. We will focus here on Macedonian and Serbian, as two languages that Franks (2019) discusses in this light. He claims that Serbian binding data is exclusive, that is, the “speakers consistently disallow coreference in comparable sentences”, whereas Macedonian speakers both allow and disallow coreference (Franks 2019: 70). We found that there is variation among speakers in both languages. Serbian speakers prefer non-binder interpretations overall, but we see some speakers who do not. The same is true for Macedonian. In other words, there is no absolute judgment but rather preference for one interpretation over the other; and the preference in both Macedonian and Serbian is for non-binder interpretations. This finding confirms that binding is very susceptible to variation.

7. Conclusion

This paper provides new experimental data to contribute to the debate on DP vs. NP and binding in South Slavic. Despić (2011) proposes that prenominal possessives in languages with DP allow coreferential readings with R-expressions or pronouns elsewhere in the clause because they are in SpecPossP position within DP, unable to c-command elements outside DP. On the other hand, prenominal possessives in languages without DP are NP-adjoined and they c-command outside their NP, disallowing coreferential readings. Such a claim predicts that Bulgarian and Macedonian, two South Slavic languages with definite articles, i.e., DP-languages, should exhibit binding potentials different from Serbian, a South Slavic language without articles, i.e., NP-language. LaTerza (2016) tests this claim and concludes that it does not hold true. Franks (2019) brings up a few interesting points regarding the data used in LaTerza 2016: the use of clitics vs full pronouns, the difference in binding potentials of nominal and prenominal possessives in Bulgarian, and the overall acceptability of such structures in Macedonian and Serbian.

This paper tests the relevant data using an online platform for distribution and recruiting native speakers who have no prior linguistic training. The results of the study show that the three languages do not exhibit differences in binding potentials, with preference for non-binder interpretations. This finding disproves the original claim that binding potentials differ between languages because of the nominal structure (DP vs. NP), as argued by the proponents of the Parameterized DP Hypothesis.

References

- Abney, Steven Paul. (1987) *The English noun phrase in its sentential aspects*. Doctoral dissertation, MIT.
- Bailyn, John Frederick. (2012) *The syntax of Russian*. Cambridge: Cambridge University Press.
- Barwise, Jon and Robert Cooper. (1981) "Generalized quantifiers and natural language". *Linguistics and philosophy* 4: 159–219.
- Bašić, Monika. (2004) *Nominal subextractions and the structure of NPs in Serbian and English*. Master's thesis, UiT The Arctic University of Norway.
- Bianchi, Valentina. (2010) "A note on backward anaphora". *Rivista di grammatica generativa – Research in generative grammar* 34: 3–34.
- Bošković, Željko. (2003) "On left branch extraction". Peter Kosta, Joanna Blaszczak, Jens Frasek, Ljudmila Geist, and Marzena Zygis, eds. *Investigations into formal Slavic linguistics: Contributions of the Fourth Conference on Formal Description on Slavic Languages – FDSL 4*. Frankfurt am Main: Peter Lang, 543–77.

- Bošković, Željko. (2005) "On the locality of left branch extraction and the structure of NP". *Studia linguistica* 59: 1–45.
- . (2008) "What will you have, DP or NP?" Emily Elfner and Martin Walkow, eds. *NELS 37: Proceedings of the 37th annual meeting of the North East Linguistic Society*. Amherst: Graduate Linguistic Student Association of the University of Massachusetts at Amherst, 101–14.
- . (2012) "Phases in NPs/DPs". Ángel L. Gallego, ed. *Phases: Developing the framework*. Berlin: Mouton de Gruyter, 343–84.
- Brame, Michael. (1982) "The head-selector theory of lexical specifications and the nonexistence of coarse categories". *Linguistic analysis* 10: 321–25.
- Caruso, Đurđica Željka. (2011) "Nominal phrases in Croatian as DPs". Koichi Otaki, Hajime Takeyasu, and Shin-ichi Tanigawa, eds. *Online Proceedings of GLOW in Asia Workshop for Young Scholars 2011*. Available at: <https://citeserx.ist.psu.edu/document?repid=rep1&type=pdf&doi=209220f5e40c8f8967b0e4789d-3c919859b9eca7#page=21>.
- Cheung, Candice Chi-Hang and Richard K. Larson. (2018) "'Make' causatives in English and Mandarin". *Lingua Sinica* 4: 4. DOI 10.1186/s40655-018-0036-9
- Chomsky, Noam. (1976) "Conditions on rules of grammar". *Linguistic analysis* 2(4): 303–51.
- . (1981) *Lectures on government and binding*. Dordrecht: Foris.
- Cinque, Guglielmo. (2005) *Comparative syntax*. Oxford: Oxford University Press.
- Corver, Norbert. (1992) "On deriving certain left branch extraction asymmetries: A case study in parametric syntax". *North East Linguistics Society* 22(1): 67–84.
- Delsing, Lars-Olof. (1993) *The internal structure of Noun Phrase in the Scandinavian languages*. Doctoral dissertation, University of Lund.
- Despić, Miloje. (2009) "On the structure of Serbo-Croatian NP: Evidence from binding". Jodi Reich, Maria Babyonyshev, and Darya Kavitskaya, eds. *Proceedings of FASL 17: The Yale meeting 2008*. Ann Arbor: Michigan Slavic Publications, 17–32.
- . (2011) *Syntax in the absence of Determiner Phrase*. Doctoral dissertation, University of Connecticut.
- . (2013) "Binding and the structure of NP in Serbo-Croatian". *Linguistic inquiry* 44: 239–70.
- Dubinsky, Stanley and Mila Tasseva-Kurktchieva. (2014) "On the NP/DP language frontier: Bulgarian as a transitional case". Talk given at the Linguistic Society of America Annual Meeting, Minneapolis, MN, 4 January 2014.
- Erteschik-Shir, Nomi. (1997) *The dynamics of focus structure*. Cambridge: Cambridge University Press.
- Franks, Steven. (2019) "Binding and phasehood in South Slavic revisited". *Studies in Polish linguistics* 14(2): 61–80.

- Fukui, Naoki. (1986) *A theory of category projection and its applications*. Doctoral dissertation, MIT.
- . (1988) "Deriving the differences between English and Japanese". *English linguistics* 5: 249–70.
- Fukui, Naoki and Margaret Speas. (1986) "Specifiers and projections". *MIT working papers in linguistics*. Vol. 8. Cambridge, MA: MIT Press, 128–72.
- Heim, Irene. (2002) "File change semantics and the familiarity theory of definiteness". Paul Portner and Barbara H. Partee, eds. *Formal semantics: The essential readings*. Oxford: Blackwell, 223–48.
- Ivšić, Kristina. (2008) *The structure of Serbian nominal phrase*. Master's thesis, University of Nova Gorica.
- Jackendoff, Ray. (1977) *X-bar syntax*. Cambridge, MA: MIT Press.
- Julien, Marit. (2002) "Determiners and word order in Scandinavian DPs". *Studia Linguistica* 56: 265–315.
- Kayne, Richard. (1994) *The antisymmetry of syntax*. Cambridge, MA: MIT Press.
- LaTerza, Ivana. (2014) *The DP category and Serbian nominal structure*. Doctoral dissertation, Stony Brook University, New York.
- . (2016) "Binding in English and South Slavic and the Parameterized DP Hypothesis". *Linguistic inquiry* 47(4): 741–53.
- Nicolova, Ruselina. (2008) *Bългарска граматика: Морфология* [Bulgarian grammar: Morphology]. Sofia: Universitetsko izdatelstvo "Sv. Kl. Ohridski".
- Pereltsvaig, Asya. (2013) "Noun Phrase structure in article-less Slavic languages: DP or not DP?" *Language and linguistics compass* 7: 201–19.
- Pesetsky, David. (1994) *Zero syntax*. Cambridge, MA: MIT Press.
- Progovac, Ljiljana. (1998) "Determiner Phrase in a language without determiners". *Journal of linguistics* 34: 165–79.
- Rappaport, Gilbert. (2001) "Extraction from nominal phrases in Polish and the theory of determiners". *Journal of Slavic linguistics* 8(1–2): 159–98.
- Reinhart, Tanya. (1986) "Center and periphery in the grammar of anaphora". Barbara Lust, ed. *Studies in the acquisition of anaphora*. Vol 1. *Defining the constraints*. Dordrecht: D. Reidel Publishing Company, 123–50.
- Runić, Jelena. (2006) "On the syntax of determiners in Serbian and Romanian". *Filološki pregled* 33: 75–93.
- Schoorlemmer, Erik. (2012) "Genitive noun complements in Serbo-Croatian: Extraction and case". Handout from presentation at FASL 21, Bloomington, IN, 11–13 May 2012.
- Schwarz, Florian. (2009) *Two types of definites in natural language*. Doctoral dissertation, University of Massachusetts, Amherst.
- Smith, Carlota S. (1964) "Determiners and relative clauses in a generative grammar of English". *Language* 40: 37–52.
- Srdanović, Sanja and Esther Rinke. (2020) "Possessive modifiers in Serbian: Coreference with clitics and strong pronouns". *Journal of Slavic linguistics* 28(2): 163–82.

- Stanković, Branimir. (2013) "DP or NP? The case of Serbian southeastern dialects and Macedonian language". Paper presented at SinFonIJa 6, University of Niš, 26–28 September 2013.
- Stjepanović, Sandra. (1998) "Extraction of adjuncts out of NPs". Paper presented at the workshop "Comparative Slavic Morphosyntax: 'The State of the Art'", Spencer, IN, 5–7 June 1998.
- Szabolcsi, Anna. (1983) "The possessor that ran away from home". *Linguistic review* 3: 89–102.
- Talić, Aida. (2013) "Extraordinary complement extraction". Handout from presentation at LSALAA 2013, Paris, 28 February 2013.
- . (2020) "Affixal-article languages and structural parallelism in Slavic and beyond". Peter Kosta and Teodora Radeva-Bork, eds. *Current developments in Slavic linguistics: Twenty years after*. Frankfurt am Main: Peter Lang.
- Trenkić, Danijela. (2004) "Definiteness in Serbian/Croatian/Bosnian and some implications for the general structure of the nominal phrase". *Lingua* 114: 1401–27.
- Vergnaud, Jean-Roger. (1974) *French relative clauses*. Doctoral dissertation, MIT.
- Williams, Edwin. (1997) "Blocking and anaphora". *Linguistic inquiry* 28: 577–628.
- Zlatić, Larisa. (1997) *The structure of the Serbian Noun Phrase*. Doctoral dissertation, University of Texas at Austin.
- Zubizarreta, Maria Luisa. (1998) *Prosody, focus and word order*. Cambridge, MA: MIT Press.

Ivana LaTerza
 Stony Brook University
 New York, NY, USA
 ivana.laterza@alumni.stonybrook.edu

Petya Osenova
 Sofia University "St. Kl. Ohridski"
 Sofia, Bulgaria
 osenova@uni-sofia.bg

Boban Karapejovski
 Blaže Koneski Faculty of Philology
 Ss. Cyril and Methodius University in Skopje
 Skopje, Macedonia
 karapejovski@flf.ukim.edu.mk

Appendix

The survey sentences from all three languages are provided below. The first line represents the sentence used for Serbian, the second one Bulgarian, and the third one Macedonian.

The survey items below are equipped with English translations. Note, however, that while the translations of the context-setting sentences and the questions are true English translations, the meaning of the target sentence (i.e., *X's NOUN V-ed him/her/it*) was what this study actually set out to determine, so translations such as 'Marija's friend pushed her' are not used below in the function of a true English translation but rather of a simplified gloss, which is why this part of the translation is set in italics. (The names used in the Serbian, Macedonian, and Bulgarian version of an example may differ. For convenience, the English translation always uses the name provided in the Serbian version.)

1. Marijin prijatelj ju je gurnuo. Koga je Marijin prijatelj gurnuo?
 Марииният приятел я бутна. Кого бутна Марииният приятел?
 Марииниот пријател ја турна. Кого турна Марииниот пријател?
'Marija's friend pushed her. Who did Marija's friend push?'
2. Marija i Jelena se igraju na dvorištu ispred kuće. Marijin prijatelj ju je gurnuo. Koga je Marijin prijatelj gurnuo?
 Мария и Елена си играят в двора пред къщата. Марииният приятел я бутна. Кого бутна Марииният приятел?
 Марија и Елена си играа во дворот пред куќата. Марииниот пријател ја турна. Кого турна Марииниот пријател?
'Marija and Jelena are playing in the yard in front of the house. Marija's friend pushed her. Who did Marija's friend push?'
3. Jelenina mačka ju je ogrebala. Koga je Jelenina mačka ogrebala?
 Еленината котка я одраска. Кого одраска Еленината котка?
 Еленината мачка ја изгреба. Кого изгреба Еленината мачка?
'Jelena's cat scratched her. Who did Jelena's cat scratch?'
4. Jelena ima mačku. Zoran je došao u posetu kod Jelene. Jelenina mačka ju je ogrebala. Koga je Jelenina mačka ogrebala?
 Елена има котка. Иван е дошъл на гости на Елена. Еленината котка я одраска. Кого одраска Еленината котка?

Елена има мачка. Зоран дојде на гости кај Елена. Еленината мачка ја изгреба. Кого изгреба Еленината мачка?

'Jelena has a cat. Zoran is visiting Jelena. *Jelena's cat scratched her.* Who did Jelena's cat scratch?'

5. Lanina mama ju je zagrlila. Koga je Lanina mama zagrlila?

Борянината мајка ја прегърна. Кого прегърна Борянината мајка?
Анината мајка ја прегърна. Кого прегърна Анината мајка?

'Lana's mother hugged her. Who did Lana's mother hug?'

6. Mila je došla kod Lane da se igraju. Lanina mama ju je zagrlila. Koga je Lanina mama zagrlila?

Мила е дошла да си играе с Боряна. Борянината мајка ја прегърна. Кого прегърна Борянината мајка?

Мила дојде кај Ана да си играат. Анината мајка ја прегърна. Кого прегърна Анината мајка?

'Mila came to play with Lana. *Lana's mother hugged her.* Who did Lana's mother hug?'

7. Lukin papagaj ga je ugrizao. Koga je Lukin papagaj ugrizao?

Ивановият папагал го ухапа. Кого ухапа Ивановият папагал?

Ивановиот папагал го гризна. Кого гризна Ивановиот папагал?

'Luka's parrot bit him. Who did Luka's parrot bite?'

8. Luka ima papagaja. Marko se igra s njim. Lukin papagaj ga je ugrizao. Koga je Lukin papagaj ugrizao?

Иван има папагал. Мартин си играе с него. Ивановият папагал го ухапа. Кого ухапа Ивановият папагал?

Иван има папагал. Марко си игра со него. Ивановиот папагал го гризна. Кого гризна Ивановиот папагал?

'Luka has a parrot. Marko is playing with him/it. *Luka's parrot bit him/it.* Who did Luka's parrot bite?'

9. Markov brat ga je udario. Koga je Markov brat udario?

Стояновиот брат го удари. Кого удари Стояновиот брат?

Марковиот брат го удри. Кого го удри Марковиот брат?

'Marko's brother hit him. Who did Marko's brother hit?'

10. Marko i Lana su se posvajali oko igračke. Markov brat ga je udario. Koga je Markov brat udario?
 Стоян и Боряна се скараха за играчките. Стояновиот брат го удари. Кого удари Стояновиот брат?
 Марко и Ана се скараа за играчките. Марковиот брат го удри. Кого го удри Марковиот брат?
 'Marko and Lana quarreled about a toy. *Marko's brother hit him.* Who did Marko's brother hit?'
11. Danilova sestra ga je poljubila. Koga je Danilova sestra poljubila?
 Пенчовата сестра го целуна. Кого целуна Пенчовата сестра?
 Зорановата сестра го бакна. Кого бакна Зорановата сестра?
 'Danilo's sister kissed him. Who did Danilo's sister kiss?'
12. Goran i Danilo su nacrtali crtež za rođendanski poklon. Danilova sestra ga je poljubila. Koga je Danilova sestra poljubila?
 Васил и Пенчо нарисоваха картичка за подарџка за рождения ден. Пенчовата сестра го целуна. Кого целуна Пенчовата сестра?
 Горан и Зоран нацртаа цртеж како роденденски подарок. Зорановата сестра го бакна. Кого бакна Зорановата сестра?
 'Goran and Danilo did a drawing as a birthday present. *Danilo's sister kissed him.* Who did Danilo's sister kiss?'
13. Njen prijatelj je gurnuo Mariju. Čiji prijatelj je gurnuo Mariju?
 Нейниот пријател бутна Марија. Чий пријател бутна Марија?
 Нејзиниот пријател ја турна Марија. Чиј пријател ја турна Марија?
 'Her friend pushed Marija. Whose friend pushed Marija?'
14. Marija i Jelena se igraju na dvorištu ispred kuće. Njen prijatelj je gurnuo Mariju. Čiji prijatelj je gurnuo Mariju?
 Марија и Елена си играат на двора пред куќата. Нейниот пријател бутна Марија. Чий пријател бутна Марија?
 Марија и Елена си играа во дворот пред куќата. Нејзиниот пријател ја турна Марија. Чиј пријател ја турна Марија?
 'Marija and Jelena are playing in the yard in front of the house. *Her friend pushed Marija.* Whose friend pushed Marija?'

15. Njena mačka je ogrebala Jelenu. Čija mačka je ogrebala Jelenu?
 Нейната котка одраска Елена. Чия котка одраска Елена?
 Нејзината мачка ја изгреба Елена. Чија мачка ја изгреба Елена?
'Her cat scratched Jelena. Whose cat scratched Jelena?'

16. Jelena ima mačku. Zoran je došao u posetu kod Jelene. Njena mačka je ogrebala Jelenu. Čija mačka je ogrebala Jelenu?
 Елена има котка. Иван е дошъл на гости на Елена. Нейната котка одраска Елена. Чия котка одраска Елена?
 Елена има мачка. Зоран дојде на гости кај Елена. Нејзината мачка ја изгреба Елена. Чија мачка ја изгреба Елена?
'Jelena has a cat. Zoran is visiting Jelena. Her cat scratched Jelena. Whose cat scratched Jelena?'

17. Njena mama je zagrlila Lanu. Čija mama je zagrlila Lanu?
 Нейната майка прегърна Боряна. Чия майка прегърна Боряна?
 Нејзината мајка ја прегрна Ана. Чија мајка ја прегрна Ана?
'Her mother hugged Lana. Whose mother hugged Lana?'

18. Mila je došla kod Lane da se igraju. Njena mama je zagrlila Lanu. Čija mama je zagrlila Lanu?
 Мила е дошла да си играе с Боряна. Нейната майка прегърна Боряна. Чия майка прегърна Боряна?
 Мила дојде кај Ана да си играат. Нејзината мајка ја прегрна Ана. Чија мајка ја прегрна Ана?
'Mila came to play with Lana. Her mother hugged Lana. Whose mother hugged Lana?'

19. Njegov papagaj je ugrizao Luku. Čiji papagaj je ugrizao Luku?
 Неговият папагал ухапа Иван. Чий папагал ухапа Иван?
 Неговиот папагал го гризна Иван. Чиј папагал го гризна Иван?
'His parrot bit Luka. Whose parrot bit Luka?'

20. Luka ima papagaja. Marko se igra s njim. Njegov papagaj je ugrizao Luku. Čiji papagaj je ugrizao Luku?
 Иван има папагал. Мартин си играе с него. Неговият папагал ухапа Иван. Чий папагал ухапа Иван?

Иван има папагал. Марко си игра со него. Неговиот папагал го гризна Иван. Чиј папагал го гризна Иван?

'Luka has a parrot. Marko is playing with him/it. *His parrot bit Luka.* Whose parrot bit Luka?'

21. Njegov brat je udario Marka. Čiji brat je udario Marka?

Неговиот брат удари Стоян. Чий брат удари Стоян?

Неговиот брат го удри Марко. Чиј брат го удри Марко?

'His brother hit Marko. Whose brother hit Marko?'

22. Marko i Lana su se posvajali oko igračke. Njegov brat je udario Marka. Čiji brat je udario Marka?

Стоян и Боряна се скараха за играчките. Неговиот брат удари Стоян. Чий брат удари Стоян?

Марко и Ана се скараа за играчките. Неговиот брат го удри Марко. Чиј брат удри Марко?

'Marko and Lana quarreled about a toy. *His brother hit Marko.* Whose brother hit Marko?'

23. Njegova sestra je poljubila Danila. Čija sestra je poljubila Danila?

Неговата сестра целуна Пенчо. Чия сестра целуна Пенчо?

Неговата сестра го бакна Зоран. Чија сестра го бакна Зоран?

'His sister kissed Danilo. Whose sister kissed Danilo?'

24. Goran i Danilo su nacrtali crtež za rođjendanski poklon. Njegova sestra je poljubila Danila. Čija sestra je poljubila Danila?

Васил и Пенчо нарисуваха картичка за подарџка за рождения ден. Неговата сестра целуна Пенчо. Чия сестра целуна Пенчо?

Горан и Зоран нацртаа цртеж како подарок за роденден. Неговата сестра го бакна Зоран. Чија сестра го бакна Зоран?

'Goran and Danilo did a drawing as a birthday present. *His sister kissed Danilo.* Whose sister kissed Danilo?'

The Syntax of Bulgarian *edin* ‘one’*

Luca Molinari

Abstract: The present paper aims at offering a syntactic model for the grammaticalization of the numeral *edin* ‘one’ in Bulgarian. *Edin* is argued to be at the beginning of the last stage of grammaticalization, i.e., the stage of the indefinite article (cf. Geist 2013). It may function as (i) a cardinal numeral; (ii) a specificity marker (individuating the referent); or (iii) an article-like element with non-referential interpretation in generic sentences. The proposal put forth here is that these different functions are the manifestation of three different structural positions: (i) the specifier of a functional projection (NumP) below the DP for the cardinal; (ii) SpecDP for the specific marker; and (iii) the head D for the article-like marker of genericity. This model represents a perfect linguistic cycle, which suggests that the present analysis may be on the right track.

1. Introduction

The present work arises from the necessity of providing a detailed description of the peculiar syntax of the numeral ‘one’ in Slavic languages, exploring here the case of Bulgarian *edin*. This need is dictated by two main reasons. The first is the fact that ‘one’ patterns differently from the other cardinal numerals (at least as far as Slavic languages are concerned). Second is the fact that ‘one’ is the only numeral which undergoes a cross-linguistically

* Parts of this paper were discussed at the following conferences: FASL30 (online, MIT, May 14, 2021), SinFonJA14 (online, University of Novi Sad, September 22, 2021), LingBaW8 (online, University of Lublin, October 15, 2021), and ISTAL25 (Aristotle University of Thessaloniki, May 15, 2022). I express my gratitude to the audiences of those events for their insightful questions, and to two anonymous reviewers for their constructive criticism. I wish to thank Paweł Rutkowski and Giuliana Giusti for their helpful comments.

I would also like to thank Iliyana Krapova for her judgments and fruitful discussion. I am grateful to Assia Assenova, Gergana Xristova, and Marija Gančeva for their judgments, and the 46 anonymous native speakers who provided me with their grammaticality judgments by completing an online questionnaire. All errors are entirely mine.

consistent process of grammaticalization, leading it to become, in the ultimate stage, an indefinite article.

As will be shown, it is not possible to map the syntax of the numeral *edin* 'one' onto one single position in the structure; thus a more detailed model is necessary to accommodate its grammaticalization process. This would lead mainly to two consequences. The most immediate one is the fact that having such a model would allow more effective probing of the diachronic data, thus tracing the path of development of the various functions of *edin* from a historical point of view. The second consequence has broader scope: building a model for the development of *edin* (and possibly extending it to other languages) would provide us with a possible explanation of the regularities of the grammaticalization path this numeral undergoes cross-linguistically. In fact, the creation of a syntactic model would suggest that the cross-linguistic regularities in grammaticalization are to some extent driven by syntax, which constrains the set of possible operations that apply to lexical items.

The aim of this paper is to fill a gap in the literature by providing a syntactic analysis for the numeral 'one' in Bulgarian. This will be achieved by showing its multifunctional nature and by designing a model that maps its various functions onto different syntactic positions in the structure.

This section provides a brief overview of the two main facts which make this analysis necessary, namely, the panorama of Slavic numerals (§1.1) and the path of grammaticalization 'one' undergoes cross-linguistically (§1.2). The rest of the paper is divided as follows: section 2 individuates the stage of grammaticalization of *edin* following Geist 2013. Section 3 introduces the reader to the theoretical framework on nominal expressions, focusing on Bulgarian. Section 4 explores the position of *edin* in its numeral function. Section 5 looks at the position of *edin* as a specificity marker, while section 6 deals with the syntax of non-specific *edin*. Section 7 is dedicated to the discussion of whether the development *edin* undergoes may be considered an instance of grammaticalization. Section 8 concludes the paper.

1.1. Panorama of Slavic Numerals

Syntactic literature has largely acknowledged the complex and entangled panorama of numerals in Slavic languages, as their behavior cannot be reduced to a single syntactic category or a single merging point in the nominal structure (cf. Franks 1994 for an overview). In languages with overt case morphology (e.g., Russian, Polish, etc.), numerals are commonly divided into different classes, according to their properties: whether they agree with the noun they quantify; whether they assign genitive case to their complement in all structural configurations; or whether they assign genitive only in structural case positions, while they agree for case in oblique configurations (for a classification of numerals in Polish, see Rutkowski 2007: 90).

Bulgarian, a language of the South Slavic group, does not display a complete paradigm of nominal case morphology¹ but has developed a fully-fledged system of definite articles. Still, Bulgarian cardinal numerals display a peculiar pattern when quantifying masculine nouns (cf. §4.1 for a more detailed discussion).

In this complex panorama, the numeral 'one' seems to have a peculiar status, as it does not pattern along with the other cardinal numbers in the languages mentioned above. In fact, 'one' in Slavic always agrees in gender, number (even displaying a plural form), and case (in languages displaying overt case morphology) with the quantified noun, never assigning it an independent case. Its full agreement paradigm is taken as evidence for its adjectival nature (cf. Giusti and Leko 2005: 145 for BCS; Ionin and Matushansky 2018: 175–6 for Russian; Rappaport 2003: 124 for Polish). Bulgarian is perfectly in line with the picture just described. Despite the impossibility of applying the case assignment diagnostics, *edin* 'one' deviates from the pattern of the other numerals (cf. §4.1).

1.2. The Path of Grammaticalization of the Numeral 'one'

As already noticed by Givón (1981), the numeral 'one' undergoes a cross-linguistically consistent process towards becoming a marker for indefinite singular nouns. This is the case for Italian and Spanish *uno*, for German *ein*, for Swedish *en*, and for Turkish *bir*. This tendency is found in Slavic languages as well: a non-exhaustive list of examples includes Upper Sorbian *jen* and Lower Sorbian *jan*, Czech *jeden* (Heine and Kuteva 2006; Caruso 2012, 2016), Molise Slavic *na* (Breu 2012), Slovene dialects of Friuli *ni* (Benacchio 2018), Macedonian *eden* (Weiss 2004), Polish *jeden* (Hwaszcz and Kędzierska 2018), and Bulgarian *edin* (Geist 2013).

This process leading to the development of an indefinite marker (and, ultimately, of an indefinite article) out of the numeral 'one' has been referred to as an instance of grammaticalization (cf. Givón 1981; Heine 1997; Heine and Kuteva 2006; van Gelderen 2011, *inter alia*). One observation is in order here: the evolution 'one' undergoes is different from other common grammaticalization processes, e.g., the well-known Jespersen cycle (cf. Jespersen 1917).² While in the latter the newly grammaticalized item substitutes the "old" one (which is then lost), in the case of 'one' the newly grammaticalized functions coexist with the older ones. However, in line with the previously mentioned

¹ As an anonymous reviewer points out, some residues of case are still visible on pronouns (e.g., *tja* 'she.NOM' vs. *neja* 'she.ACC') and clitics (e.g., *go* 'he.ACC' vs. *mu* 'he.DAT').

² I thank an anonymous reviewer for having pointed out this issue.

authors, I presuppose that this process is governed by the same principles that govern grammaticalization (cf. §7.3 for a justification of this view).³

The path of grammaticalization of the numeral 'one' follows some consistent stages and goes along with a process of semantic bleaching, in which the numeral assumes the functions of an indefiniteness marker at the expense of its quantificational nature (cf. Givón 1981).

Givón (1981: 50) provides a tripartite model of grammaticalization of the development of the functions carried out by the numeral 'one'. His model is reported here in (1).

- (1) Quantification > referentiality/denotation > genericity/connotation

Heine (1997: 72–74) offers instead a more detailed picture of this diachronic change, subdividing it into five different stages, summarized in (2):

- (2) Stages of grammaticalization of the numeral 'one':
- I. The numeral: 'one' has only a quantitative function (corresponding to Givón's first stage).
 - II. The presentative marker: 'one' introduces a new salient referent (i.e., expected to be taken up in subsequent discourse), which is supposed to be unknown to the hearer.
 - III. The specific marker: 'one' introduces referents which are known to the speaker, but presumed to be unknown to the hearer, independently of their saliency (II and III correspond to Givón's second stage).
 - IV. The non-specific marker: 'one' introduces a referent which is unknown both to the speaker and to the hearer and whose reference is not important in the discourse.
 - V. The generalized article: the article can occur with almost all nominal classes, and its insertion is justified by mere syntactic reasons (IV and V equal Givón's last stage).

³ Various authors (cf. Diewald 2011 and references therein for an overview) argued in favor of distinguishing the process of grammaticalization from that of *subjectification* (cf. Traugott 1989) and *pragmaticalization* (cf. Aijmer 1997). While subjectification is an instance of semantic change which does not contradict the nature of grammaticalization and goes hand in hand with it, pragmaticalization (which mainly applies to discourse markers) arises from the need to keep the domain of "grammar" and that of "pragmatics" separated (Diewald 2011: 384). However, as Diewald argues, a richer notion of "grammar" that encompasses pragmatic functions allows us to treat pragmaticalization as another instance of grammaticalization.

The grammaticalization of 'one' is generally monodirectional (from stage I to V)⁴ and cumulative: once the numeral has reached a given stage, it must have acquired all the functions proper to all the preceding ones. There is, however, a certain degree of overlap among adjacent stages.

Given these premises, let us turn to the development of the numeral *edin*⁵ 'one' in Bulgarian, which will allow us to map its different syntactic positions in the nominal expression.

2. Evaluating the Stages of Bulgarian *edin*

Geist (2013) adopts Givón's and Heine's grammaticalization models to evaluate the stage of development of *edin* in Bulgarian.

The initial stage of the process is that of the numeral, in which "'one' expresses the fact that the set to which the referent of the noun phrase belongs consists of exactly one entity" (Schroeder 2006: 556). To distinguish this function from other ones that are carried out by *edin*, Geist provides some tests that single this interpretation out, such as the modification by particles emphasizing its cardinality, shown in (3), and the possibility of being contrasted with other numerals, as in (4).

- (3) Samo *edin* telefon li imate?
 only one.M telephone Q have.2PL⁶
 'Do you have only one telephone (or two)?'

- (4) Ivan ima *edin* sin (, a ne dvama).
 Ivan has one.M son and not two

'Ivan has one son (and not two).' (Geist 2013: 127–28)

The next stage is that of the presentative marker, in which 'one' is used only with referential noun phrases (NPs) that the speaker wants to mark as salient in the discourse. The salience of the referent can be expressed by taking up the NP in subsequent discourse. In the case of Bulgarian, the NP introduced

⁴ But see Joseph 2011 and Trousdale and Norde 2013 for counterexamples to the unidirectionality.

⁵ I use the unmarked masculine form *edin* to refer to all singular forms of the numeral, i.e., *edna* 'one.F' and *edno* 'one.N'. The plural form of 'one' will be referred to as simply *edni*.

⁶ The following abbreviations are used throughout the paper: M = masculine, F = feminine, N = neuter, 1 = 1st person, 2 = 2nd person, Q = polar question particle, ACC = accusative, DAT = dative, REFL = reflexive particle, DEF = definite article, SG = singular, PL = plural, BF = *brojna forma* 'count form', COND = conditional, HUM = human, IND = indicative, SUBJ = subjunctive.

by *edin* can be referred back to in subsequent discourse—e.g., by a personal pronoun, as in (5)—but does not need to, as in (6).

- (5) Imalo edno vreme *edin* starec. Toj imal trima sina.
 had one.N time one.M old.man he had three sons
 ‘Once upon a time, there was an old man. He had three sons.’
- (6) Predi da predam statijata ja dadox na *edin*
 before *da* submit.1SG paper.DEF it.ACC gave.1SG to one.M
kolega za korekcii. Sled tova podadox statijata
 colleague for corrections after that submitted.1SG paper.DEF
 na edno spisanie.
 to one.N journal
 ‘Before submitting my paper, I gave it to a colleague for proofreading.
 Then I sent the paper to a journal.’ (Geist 2013: 131)

The third stage is named by Heine that of the “specific marker”, referring to the fact that ‘one’ can denote a referent or an entity which is known to the speaker and new to the hearer. Geist (2013) points out that *edin* needs to satisfy the condition of *identifiability* in the sense of Ionin 2013 (esp. p. 82), i.e., the speaker should be able to answer the question “which X is it?”. This is shown in (7).

- (7) a. Čete mi se *edno* spisanie.
 read I.DAT REFL one.N journal
 ‘I would like to read a journal.’
- b. A imenno, poslednijat broj na Novo Vreme.
 and namely last.DEF issue of Novo Vreme
 ‘Namely the last issue of Novo Vreme.’
- c. #Kakvoto i da e.
 which and *da* be
 ‘Any journal would do.’ (Geist 2013: 132)

The last two stages in Heine’s path of grammaticalization are conflated by Geist into the stage of the “indefinite article”, à la Givón. This stage entails an obligatoriness condition: since fully-fledged articles are pure syntactic markers, they are inserted in the structure for mere syntactic requirements. Bulgarian does not completely meet this requirement, as the appearance of *edin* is not generally obligatory, as shown in (8) (Ivanova and Koval’ 1994: 59, cited in Geist 2013: 136). The only instance of obligatory appearance of

this marker is with bare NPs in topic position (to mark the “aboutness topic”, which needs to be specific), as in (9) (from Ivančev 1957: 515, cited in Friedman 1976: 338).

- (8) V stajata vlezete.
in room.DEF came child
'A child came into the room.'
- (9) a. Edna žena ja risuva edin xudožnik.
one.F woman she.ACC painted one.M painter
'A woman was painted by a painter.'
- b. Ženata ja risuva edin xudožnik.
woman.DEF she.ACC painted one.M painter
'The woman was painted by a painter.'
- c. *Žena ja risuva edin xudožnik.
woman she.ACC painted one.M painter
Intended: 'A woman was painted by a painter.'

Moreover, Geist identifies three further requirements that the indefinite article should meet: (i) it should be used non-referentially in generic contexts; (ii) it should be able to occur in predicative position in combination with predicative nouns; and (iii) it should have non-referential use in modal and negative scope. Of these three outlined features, only the first one is displayed by *edin*, (10) (even though, as noted by Geist, in some instances it can be omitted), while the latter two are not met, (11–13).

- (10) *(*Edin*) džentâlmen vinagi otvarja vrata na damite.
one.M gentleman always opens doors to ladies.DEF
'A gentleman always opens doors for ladies.' (Geist 2013: 142)
- (11) Peter e (**edin*) učitel.⁷
Peter is one.M teacher
'Peter is a teacher.'
(Ivanova and Koval' 1994: 59, cited in Geist 2013: 139)

⁷ An anonymous reviewer points out that in (11) the presence of *edin* does not make the sentence ungrammatical, but it could only answer the question “Who is Peter?”, and not “What does Peter do?”. Thus, *edin* can only be used in identificational copular sentences (but not in predicational ones) with specific indefinite reference (cf. also Geist 2013).

- (12) a. Tja iska da se omâži za edin rusnak.
 she wants *da* REFL marry to one.M Russian
 ‘She wants to marry a certain Russian man.’ (Geist 2013: 143)
- b. Continuation compatible with (12a): I know him.
- c. Continuation not compatible with (12a): #There are no candidates yet.
- (13) Toj ne spomena edna podrobnost.
 he not mentioned one.F detail
 ‘He didn’t mention some detail.’ (Geist 2013: 144)

The stage of Bulgarian *edin* resulting from the different diagnostics is summarized in the schema in (14) (adapted from Geist 2013: 147).

- (14) Stages and functions of markers of indefinite reference
- | | | |
|--|-----------------------------|-----------------------|
| i. The numeral |] I. Numeral | |
| ii. The presentative marker | | |
| iii. The specificity marker |] II. Indefinite determiner | |
| iv. Predicative use / generic use | | |
| v. Non-referential use in modal and negative scope |] III. Indefinite article | Bulgarian <i>edin</i> |
| vi. The generalized article | | |

Given the evidence above, Geist concludes that the stage of development of Bulgarian *edin* could be roughly placed at the beginning of the stage of the indefinite article, as it displays at least some of the features that are typical of this last stage (though it has not reached the status of a full-fledged indefinite article yet).

3. The Theoretical Framework

Now that the different functions *edin* may carry out have been presented, it is necessary to set some theoretical assumptions that constitute the starting point of the analysis of the complex syntax of *edin*. Section 3.1 deals with the general analysis of nominal expressions in the relevant framework assumed here. Section 3.2 focuses instead on the structure of Bulgarian nominal expressions.

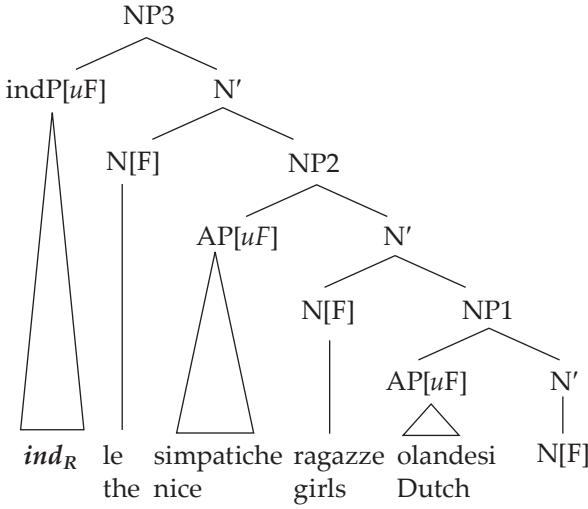
3.1. The Structure of Nominal Expressions

The structure of the nominal expression assumed here follows the line of inquiry started by Giusti (1994, and subsequent works). Giusti distinguishes the elements that were previously grouped under the label “determiners” (i.e., articles, demonstratives, and quantifiers) into different classes occupying different positions inside the extended nominal projection.

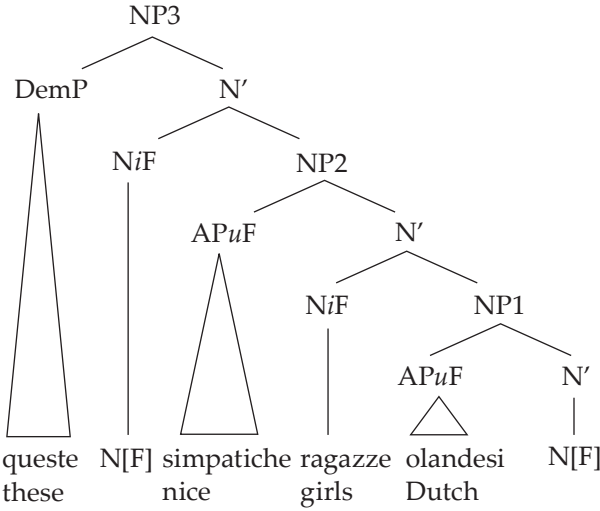
Giusti (2002) formalizes an organic analysis of nominal expressions, adopting the idea that the functional layers above the NP are extended projections—in the sense of Grimshaw 1991—of the head *N*(oun). The main idea is that *N* reprojects as many times as necessary to satisfy its Selection (theta-role assignment) and Modification (combination with adjectives) requirements. Each time *N* reprojects, a new functional head containing silent functional φ -features of *N* is created, which in turn allows for the creation of an empty specifier slot which can host adjectival modifiers (APs). The APs merged in the empty specifiers (as proposed by Cinque 1994) share with the head *N* number, gender, and case features via Concord (Giusti 2008). Concord is instantiated as the modifier is first merged in the specifier of a functional projection (functional projections in the inflectional layer are assumed not to be labeled for any feature but are mere copies of the φ -features of *N*) and is enhanced by the Spec-Head configuration. This mechanism is always local and does not trigger any movement. The highest reprojection of *N* may be labeled *D*(eterminer)*P*(hrase).⁸ Its specifier hosts referential elements like demonstratives, possessive pronouns, and proper nouns. These are all maximal projections that raise to Spec*DP* to have their referential feature checked, as the interpretation of the whole nominal expression is assumed to take place in that position at LF (Giusti 2002: 106). Given that the interpretation takes place in the left edge, articles are just dummies, heads which spell out the functional features of the nominal expression in the highest functional head, i.e., *D*. The overt realization of *D* licenses an empty operator in Spec*DP* (in the spirit of Campbell 1996), responsible for the interpretation of the whole nominal expression. A schematic representation of a nominal expression realizing the article is given in (15a), in which the highest reprojection is labeled “NP3” and the silent operator licensed by the definite article is indicated as “ind(exical)_R”. In the case in which a referential element appears directly at the left edge, there is no need for any silent operator, and the highest head (labeled *N* in the examples in (15b)) can remain covert for Economy reasons, as the features are already retrievable from the specifier.

⁸ Giusti labels this projection in different ways, e.g., *FP*^{max} in Giusti 2002, *NP*_{*n*} in Giusti 2011. Independently of the label, the highest projection corresponds to what is traditionally referred to as *DP*, so I am using this label for ease of exposure.

(15) a.



b.



(Giusti 2011: 115-6)

The realization of the article is strictly related, even from a diachronic point of view, to the realization of morphological case (cf. Giusti 1995). This is a piece of evidence supporting Giusti's (1994, 2002, 2008, 2015) conclusion that D is the *locus* in which Case is assigned, and articles are just a bundle of case, number, and gender features of the nominal expression⁹ (but cf. §3.2 for Bulgarian).

⁹ This claim is strengthened by the distribution of nominal case morphology in languages that also display articles, as is the case of German. In this language, the

3.2. Bulgarian Nominal Expressions

Capitalizing on the leading principle of this theoretical framework outlined in §3.1, Giusti and Dimitrova-Vulchanova (1996; Dimitrova-Vulchanova and Giusti 1998; henceforth, D-V&G) build a syntactic analysis of the nominal expression in Bulgarian. They argue that, in Bulgarian, N never moves to D in overt syntax: if this were the case, one would expect a sentence like (16a) to be possible, contrary to the facts. The grammatical alternative is instead given in (16b). When N does not display any modifier, the enclitic article attaches directly to it, as in (16c).

- (16) a. *momče-to goljamo
 boy-DEF big
 Intended: 'the big boy'

- b. goljamo-to momče
 big-DEF boy
 'the big boy'

- c. momče-to
 boy-DEF
 'the boy'

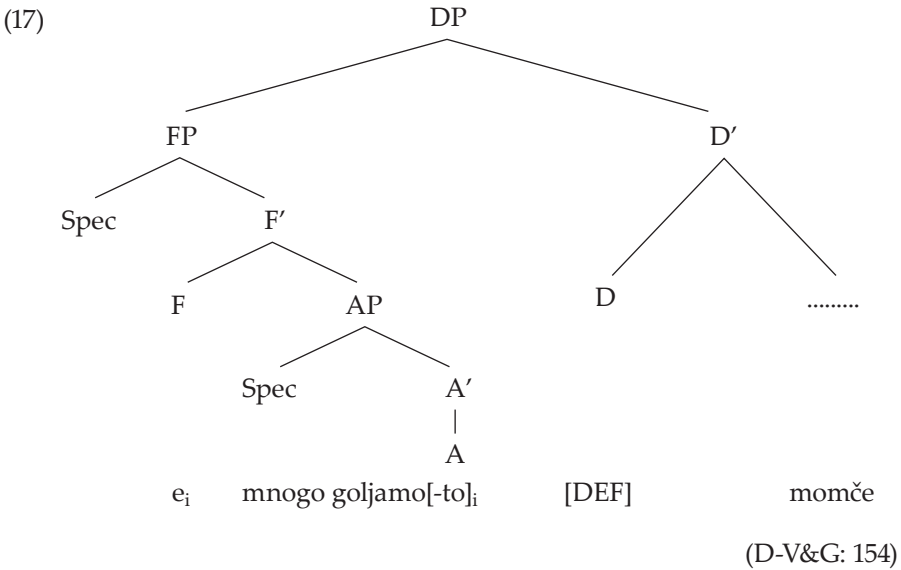
(D-V&G: 149)

One of the key assumptions is that the so-called "definite article" in Bulgarian is not an element merged in D, but it is a form of "definiteness" inflection which originates and is checked in a functional projection FP projected by the noun or the nominal modifier itself (Giusti and Dimitrova-Vulchanova 1996: 126). The head D is instead specified for a [DEF(INITENESS)] feature with which the definite article agrees in a local configuration to be interpreted. Thus, "bare" Ns procrastinate the movement to D at LF to check the features of the definite article (D-V&G: 149). When adjectival modifiers are present instead (e.g., in (16b)), the higher AP whose head is inflected for the definite article raises to SpecDP to allow for feature-checking.¹⁰ The scenario depicted so far is nicely represented in (17) on the following page, which is the structure

"strongest" case morpheme appears on the article (when it is present), while case morphemes on the other elements in the nominal expression are a form of agreement with the relevant functional head (i.e., the one spelled out by the article) (cf. Giusti 1995).

¹⁰ The movement of APs with definite inflection to SpecDP also brings evidence in favor of the analysis of the Bulgarian definite article not being merged in D. If the definite article in Bulgarian arose in D, one would expect it to appear at the right of complex adjectival modifiers moving as a unique block in SpecDP, as in (i). Contrary to the facts, (i) is ungrammatical. The structure in (ii) is what one gets instead. The

of the expression in (16b). Note that the position of the AP is derived via movement to SpecDP.



This movement is necessary to satisfy a generic principle formulated by D-V&G (p. 158) and referred to as the “Doubly Filled XP Filter”, reported here in (18).

- (18) A functional projection must be visible at all levels of representation by either
- a. making the specifier visible, and/or
 - b. making the head visible.

The conditions in (a) and (b) may be applied conjointly or disjointly, depending on the language-specific parametric setting of this filter. In Bulgarian, the conditions are disjoint: either the specifier or the head of a functional projection can be spelled out, but both cannot be realized at the same time. AP-to-DP

example in (ii) perfectly fits the account in which the article is generated as the head of a functional projection FP projected by the AP itself, as represented by the brackets.

- (i) *_{[AP (Mnogo) [_{A'} [_{A°} veren] [_{PP} na žena si]]]} -jat mǎž.
 very true to wife his -the man
- (ii) _{[AP (Mnogo) [_{A'} [_{A°} vern-i] -jat] [_{PP} na žena si]} mǎž.
 very true -the to wife his man

‘The man very true to his wife.’

(D-V&G: 156)

movement of an “inflected” adjective makes the functional projection visible, as it spells out the relevant nominal features. In this way, the “inflected” AP in SpecDP makes the nominal φ -features contained in D retrievable; thus, the head D need not (for Economy reasons, must not) be overtly spelled out.

The filter in (18) accounts for structures such as (19) which display a kind of “double definiteness” that is ruled out in Bulgarian.¹¹

- (19) **verni-jat mǎž-ât*
 true-DEF man-DEF

Intended: ‘the faithful man’

(D-V&G: 157)

According to what has been said so far, the adjective *verni-jat* in (19) raises to SpecDP, hence making the functional projection visible at the interface. Being the highest nominal element in the nominal expression, it has precedence in this upwards movement. Furthermore, the definite article appearing on the head N *mǎž-ât* requires it to covertly raise to D in order for the article to be interpreted. This creates a condition in which both the specifier and the head of the DP would be filled. This clashes with the filter in (18), which applies disjointly and rules out the structures.

The line of inquiry briefly presented contains three key points that will be crucial for the following discussion: (i) the fact that demonstratives are found in SpecDP, a position in which their referential features are checked and in which the interpretation of the whole nominal expression takes place; (ii) the fact that the position D is associated with the abstract representation of the nominal φ -features of N; and (iii) the fact that in Bulgarian the filter in (18) is interpreted disjointly, i.e., if the specifier of a functional projection is visible, the head need not (therefore must not) be overtly realized (and the other way around).

¹¹ As pointed out by an anonymous reviewer, Bulgarian does display instances of double definiteness, defined as “Multiple Determination” (cf. Rudin 2019; Franks 2020). These are structures in which the demonstrative may co-occur with the definite inflection on adjectives, as in (i), but never on nouns, shown in (ii). It may be argued that in cases like (i), the AP does not move to SpecDP in that the position is already taken by the demonstrative. Thus, a construction such as the one in (ii) is ruled out in the same way (19) is. The filter in (18) is thus compatible with the phenomenon of MD.

- (i) *Ax, tezi tvoi-te krasivi oči!*
 ah these your-DEF beautiful eyes
 ‘Ah, those beautiful eyes of yours!’

- (ii) **tazi tetradka-ta*
 this notebook-DEF
 Intended: ‘this notebook’

4. *Edin* as a Cardinal Numeral

It was shown above that the primary function *edin* carries out is quantificational. The superficial form of this numeral, however, resembles an adjective, as it agrees for φ -features with the noun it occurs with. The agreement pattern it displays is a crucial indicator of its structural position, which will be explored in what follows. Section 4.1 underlines the differences between *edin* and the other cardinal numerals and argues for a specifier status of the former. Section 4.2 tries to support the specifier status of *edin*, bringing evidence from the existence of a specific idiomatic expression.

4.1. *Edin* vs. the Other Cardinal Numerals

There is strong agreement in the literature about the fact that the numeral ‘one’ in Slavic languages indeed has an adjectival nature: in languages that have overt case morphology, the numeral ‘one’ never assigns case, but only agrees with the quantified noun (see, for example, Franks 1994: 650, 664). Moreover, the *Mittelfeld* position of numerals is well acknowledged: as Cinque (2005) showed, in all the possible orderings of nominal modifiers, numerals are always internal to the nominal expression (i.e., they are structurally lower than demonstratives). This pattern led to the idea that ‘one’ is merged in the specifier position of a functional projection which is lower than the DP (cf. Rutkowski 2007). As a result, the specifier enters in a Spec-Head agreement relation with the head noun and agrees with it.

Bulgarian does not have overt case but patterns in line with the other Slavic languages in that the numeral *edin* agrees for φ -features with the NP it quantifies. This is in line with the assumption that cardinal *edin* is merged in the specifier position of a DP-internal functional projection in the nominal spine. For simplicity, I will adopt here Giusti and Dimitrova-Vulchanova’s (1996) label “NumP” for this projection hosting numerals.

Bulgarian is in line with the general pattern in that *edin* can be preceded by other nominal modifiers traditionally associated with the DP-layer, such as demonstratives. While it is much more common to find examples with demonstratives preceding numerals greater than ‘one’, combination with *edin* is not an exception, as shown in (20) by Dimitrova-Vulchanova and Tomić (2009: 9) and in (21), taken from the Bulgarian National Corpus (<http://search.dcl.bas.bg/>; the specific file name is provided in parentheses).¹²

¹² A simple corpus search on Sketch Engine (Kilgariff et al. 2004, 2014) on the “Bulgarian National Corpus with web” (corpus size: 419,512,059 tokens) reveals this difference: the combination *tezi dva* ‘these two’ registers 10,683 occurrences (about 0.002% of the whole corpus), while *tazi edna* ‘this.F one.F’ has 117 occurrences (0.00002%).

- This difference is intuitively justified by the fact that the plurality expressed by *tezi* 'these' is always vague and thus can be specified by adding the exact number of items. The singularity of *tozi* 'this' is instead always specified and may only be spelled out by the numeral 'one'. Thus, the latter is generally omitted, unless some discourse-specific pragmatic reasons require its spell-out.

- (23) a. dva(ta) stola
 two(DEF) chair.BF
 '(the) two chairs'
- b. mnogo stolove
 many chair.PL
 'many chairs'

Another idiosyncrasy involving low cardinals, mainly (but not restricted to) 'two' to 'six' (cf. Pancheva 2018: 205), is their property of taking the suffix *-(i)ma* when co-occurring with a masculine personal noun, which appears in its regular plural form, as shown in (24).

- (24) dvama(ta) / trima(ta) mâže
 two.M.HUM(DEF) three.M.HUM(DEF) man.PL
 '(the) two/three men' (Giusti and Dimitrova-Vulchanova 1996: 142)

Edin differs from other cardinals in that it does not trigger the *brojna forma* on the NP it quantifies, (25a), as it can only be combined with singular nouns. Moreover, *edin* never takes the *-(i)ma* suffix in front of human masculine nouns, (25b). Interestingly enough, the same holds for the plural form *edni*, (26), which, despite its being plural, cannot combine with the *brojna forma* but requires the canonical plural form.

- (25) a. *edin* stol / *stola
 one.M chair.SG chair.BF
 'a chair'
- b. **edinima* mâž
 one man
 Intended: 'a man'
- (26) a. *edni* stolove / *stola¹³
 one.PL chair.PL chair.BF
 'some chairs'

¹³ The example in (26a) shows that plural *edin* patterns along with quantifiers such as *mnogo* 'much/many' (cf. example (22b)), while cardinal numbers display the same behavior as the quantifier *njakolko* 'some', which requires masculine nouns to feature the *brojna forma* (cf. Franks 2018).

- (26) b. **ednima* *mâže*
 one.PL men

Intended: 'some men'

Another difference lies in that the cardinal 'one' cannot appear after other adjectival modifiers. This is shown in sentence (27), which the consulted informants strongly rejected. It is interesting to note that the same effect is not obtained with the other cardinal numerals, (28). This may be due to the specifier position of 'one', which prevents APs from crossing it while moving upwards.

- (27) **Nova edna kniga veče e po knižarnicite*.¹⁴
 new one.F book already is at libraries.DEF

Intended: 'One new book is already available in libraries.'

- (28) a. *dve (novi) knigi*
 two new books
 'two new books'
- b. *dvete (novi) knigi*
 two.DEF new books
 'the two new books'
- c. *novite dve knigi*¹⁵
 new.DEF two books
 'the two new books'

(Giusti and Dimitrova-Vulchanova 1996: 133)

Giusti and Dimitrova-Vulchanova (1996) assume the headedness of cardinal numerals on the basis of their property of selecting the *brojna forma* and their possibility of being crossed by an AP. Since the position of the other numerals is not relevant for the present discussion, I will remain agnostic here as to the status of cardinal numbers different from *edin*. However, the agreeing nature of *edin*, its impossibility to trigger any special form on masculine nouns, and its impossibility of being crossed by an AP targeting a higher position seem

¹⁴ As was pointed out to me by Assia Assenova, if the article appeared on *nova*, i.e., *novata edna kniga*, the result would still be impossible, as the "definiteness" of the adjective sharply contrasts with the "indefiniteness" of the cardinal.

¹⁵ As an anonymous reviewer pointed out, both the order in (28b) and (28c) may be base-generated, in which case, the argumentation proposed here does not stand up. Such possibility cannot be excluded. In that case, however, the fact that the order of (28c) is not possible with *edin* would be left without an explanation.

to suggest that this cardinal is merged in a specifier position below the DP, labeled here as SpecNumP. In that structural configuration, *edin* would not be able to select any special form on the NP, and it would block AP movement, as the specifier it occupies is not available as an intermediate landing site.

4.2. *Edin* as a Specifier

A further piece of evidence for the maximal projection status of *edin* comes from the idiomatic expression *edin i sâšt*, corresponding to the English ‘one and the same’ (literally, ‘one and same’). Like its English equivalent, the expression *edin i sâšt* is used to refer to a single referent which, in the case of (29), is shared by the two subjects.

- (29) Petâr i brat mu ispolzvat *edin* i sâšt kompjutâr.
 Petâr and brother he.DAT use one.M and same computer
 ‘Petâr and his brother use the same computer.’

In this idiomatic expression, *edin* is a real cardinal numeral, as it refers to the quantity ‘one’. In fact, the NP modified by *edin i sâšt* needs to be unique; this is evident from the impossibility of using this expression with referents that cannot be shared by more than one subject, as is the case in (30).

- (30) *Dvete sestri imat *edin* i sâšt nos.¹⁶
 two.DEF sisters have one.M and same nose
 Intended: ‘The two sisters have the same nose.’

The example in (30) is ungrammatical, as it would imply that the two sisters share the same nose, which is not possible in our world. Thus, the cardinality of ‘one’ is a meaningful part of the semantics of the idiomatic expression. The sentence would be felicitous substituting *edin i sâšt* with *sâštija* ‘the same’ (lit. ‘same-the’), which in this case implies ‘of the same shape’.

If we consider idiomatic expressions to be “frozen bits of complex syntax” (Nattinger 1980: 337), we can observe that this structure is a coordination between the numeral *edin* and the adjective *sâšt*.¹⁷ On the assumption that APs are maximal projections merged as specifiers of the nominal spine, we can conclude that this structure is a coordination of two maximal projections. A tentative sketch of the structure is given in (31).

¹⁶ I thank Iliyana Krapova for this insight and for the judgment.

¹⁷ Note that the same structure is not possible with other cardinal numerals.

- (31) Petâr i brat mu ispolzvat [_{NumP} [_{&P} [_{QP} *edin*]
 Petâr and brother he.DAT use one.M
 i [_{AP} sâšt]] kompjutâr.
 and same computer

The existence of such idiomatic expression strengthens the hypothesis of having the cardinal numeral *edin* merged as a specifier, as it can be coordinated with another maximal projection. The functional projection hosting the numeral *edin* is labeled here as “NumP”. Let us now turn to the syntactic position of *edin* as a specificity marker, which constituted the intermediate stage (in Givón’s model) of grammaticalization.

5. *Edin* as an Indefinite Marker

In this section, I will deal with the status of *edin* as a specificity marker, as it has already completed the third stage of development in Heine’s scale. In this analysis, I take the stage of the “presentative marker” with that of the “specific marker” to spell out the same syntactic position, as they function in an analogous way: both are subject to Ionin’s (2006) *noteworthiness* condition (cf. §5.2.3). I will thus refer to both functions as just “specific *edin*”. Section 5.1 argues that specific *edin* occupies a higher position than that of the numeral. Section 5.2 characterizes this position as SpecDP, operating in parallel with pronouns and demonstratives.

5.1. Higher Structural Position

Before turning to the claim that the indefinite marker *edin* occurs in the DP-layer, we can verify whether this function implies that *edin* occupies a higher structural position than the other numerals. A piece of evidence in this direction comes from the behavior of the plural form *edni*, which displays an interesting pattern.

By nature, plural *edni* can be considered a real numeral only when quantifying pluralia tantum nouns, as in example (32), reported by Nicolova (2017: 194) from Maslov (1982: 367).

- (32) Tja vze samo *edni* čorapi.
 she took only one.PL socks
 ‘She took only one pair of socks.’

This plural form is, however, not restricted to pluralia tantum nouns, as it also co-occurs with plural count nouns. In the latter case, however, it would be impossible for *edni* to function as a numeral, thus it is considered an indefinite

pronoun (Nicolova 2017: 194). NPs introduced by *edni* refer to “unilaterally defined, unilaterally identifiable multitudes of phenomena named by the lexical root of the noun. The combination of the features *unilateral identifiability* + *multitude* explicates the feature *part of the whole that the speaker can identify*” (Marovska 2017: 20, my translation),¹⁸ as shown in (33).

- (33) a. *Edni* studenti zaminaxa.
 one.PL students left
 ‘Some students left.’
- b. *Edni* studenti ot vašata grupa zaminaxa.
 one.PL students from your.DEF group left
 ‘Some students of your group left.’ (Marovska 2017: 20)

Interestingly enough, *edni* can precede numerals quantifying over plural count nouns. This indicates that *edni* occurs in a structural position that is higher than the position in which numerals are merged. Compare the examples in (34), taken from Sketch Engine (Kilgarrieff et al. 2004, 2014), with the name of the corpus specified in parentheses:

- (34) a. ... a tuk idejata e *edni* dvama duši sami da
 but here idea.DEF is one.PL two souls by.themselves *da*
 precejat kakvi tajni da razkrijat...
 judge which secrets *da* reveal
 ‘...but here the idea is for two (specific) people to decide by
 themselves which secrets to reveal...’ (bgTenTen12, 200312948)
- b. No imaše *edni* dva slučaja, kogato ne uspjax.
 but had one.PL two cases when not succeed
 ‘But there were two (specific) cases in which I failed.’
 (bgTenTen12, 62821614)

Note that in the cases above, *edni* triggers a specific reading (as the speaker can identify the set containing the instances described by the lexical noun) and thus qualifies as a specific indefinite marker. The singular form is, however,

¹⁸ In Marovska’s (2017) terms, the “unilateral identifiability” corresponds to specific indefiniteness, in which the referent is identifiable by the speaker but not by the hearer (in this sense, the identifiability is unilateral).

banned in these contexts, as it would lead to a mismatch in terms of number features.¹⁹

5.2. *Edin* in SpecDP

Now that it has been shown that *edin* may occur in a structural position above the numerals, it is necessary to better understand the exact position it occupies. I will argue that *edin* occurring as a specificity marker is merged in the specifier of the highest nominal projection, namely SpecDP.

This section (and the following ones) provides examples taken from an online pilot questionnaire (using the platform Google Forms)²⁰ created for the purpose of collecting grammaticality judgments for a set of 38 sentences to test different functions of *edin*. The questionnaire was completed by 46 anonymous native speakers, who were given a Likert scale task with a range from 1 (totally ungrammatical) to 7 (perfectly grammatical).²¹

5.2.1. *Edin* and Pronouns

As a first piece of evidence, *edin* may covary with an indefinite pronoun, such as *njakakâv* (cf. (35) from Stoevski 2019: 193).²²

¹⁹ Note that it is possible to find the singular neuter form *edno* preceding another numeral, as in (i). In this case, however, we are dealing with a different function, in that *edno* indicates an approximative quantity (Stoevski 2019).

(i) Sigurno ima *edno* dvadeset godini otkak ne sâm hodil na more.
surely has one.N twenty years since not am went to seaside

'It must have been some twenty years since I last went to the seaside.'

(Stoevski 2019: 201)

²⁰ The reader can find the questionnaire at the following DOI: 10.17605/OSF.IO/AKUXF.

²¹ Reporting the mean score x of the sentences, the following notational criterion is adopted: (i) if $5 \leq x \leq 7$, the sentence is considered grammatical; (ii) if $4 \leq x < 5$, the sentence is weird, thus marked with "?"; (iii) if $1 \leq x < 4$, the sentence is ungrammatical, thus marked with an asterisk (*).

²² As noted by Stoevski (2019: 193), there are contexts in which the substitution of *edin* with an indefinite pronoun is not possible; these are contexts in which the speaker has no direct experience of the relevant referent, as is the case in a sentence like *Njakoj/*Edin čovek maj se e opitval da razbie ključalkata* 'It seems someone has been tampering with the lock.' This is in line with the specificity of *edin* as an indefinite marker and is another piece of evidence that the distribution of *edin* obeys a noteworthiness condition (cf. §5.2.3).

- (35) a. *Edna* / *njakakva* *kotka* *se* *peče* *na* *slânce* *vârhu*
 one.F some cat REFL sunbathes on sun on.top
pokriva *na* *kolata* *ti*.
 roof of car.DEF you.DAT
 'A/Some cat is sunbathing on the roof of your car.'
- b. *Na* *pokriva* *na* *kolata* *imalo* *edna* / *njakakva* *kotka*.
 on roof of car.DEF had one.F some cat
 'On the roof of the car, there was a/some cat.'

Following Giusti's (2002: 109) assumption that, among the elements that used to fall under the category of *determiners* (e.g., articles, demonstratives, quantifiers, pronouns), only articles are functional heads, the fact that specific indefinite *edin* may occupy SpecDP is automatically borne out. This does not seem to be such an unnatural assumption, as SpecDP is the place where referential elements (e.g., demonstratives and pronouns) are argued to sit.

From a structural point of view, specific *edin* patterns along with pronouns (sitting in the SpecDP) and differently from articles (which may be identified as residing in D). Pronouns are, in fact, able to license a null nominal and may be found isolated from the head N, as in (36a), while articles are ungrammatical if they do not co-occur with the N they modify (cf. (37), which is an example from Italian that has proclitic definite articles). *Edin*, not surprisingly, behaves like indefinite pronouns, (36b).

- (36) a. *Vidjah* *njako* (*njakakâv* *čovjek*) *da* *vliza* *v* *kâštata*
 saw.1SG someone some man *da* enters in house.DEF
i *se* *obadih* *na* *policijata*.
 and REFL phoned.1SG to police.DEF
 'I saw someone/some man enter the house and I called the police.'
 (Stoevski 2019: 194)
- b. *Dojdoha* *edni* (*hora*) *za* *malko* *i* *mi* *zagubiha*
 came one.PL people for little and I.DAT wasted
tri *časa*.
 three hours
 'Some folks popped in "for a moment" and wasted three hours of my time.'
 (Stoevski 2019: 187)

- (37) Ho visto il / un *(ragazzo).
 have.1SG seen the a boy (Giusti 1997: 103)

Not only is *edin* substitutable with an indefinite pronoun, but the two are generally in complementary distribution;²³ this points at the fact that the two forms compete for the same position.

5.2.2. *Edin* and Demonstratives

Taking the indefinite marker *edin* to occur in SpecDP leads to the assumption that it is found in the same position in which demonstratives are generally found (or interpreted). Let us consider Brugé's (2002) account of demonstratives in Romance. Bringing evidence mainly from Spanish, Brugé convincingly argues that demonstratives are generated low in the structure, as in this language they can appear post-nominally, following all classes of adjectives, even the lowest ones. This leads the author to conclude that demonstratives are generated in a projection which is immediately above that of the NP (before N movement takes place) but lower than the "inflectional layer" where adjectives are merged. Moreover, she claims that these items are specified for [+referential] and [+deictic] features. The fact that demonstratives are found as the highest element in the extended nominal projection in many languages leads to the assumption that they undergo movement targeting the highest nominal layer. More specifically, their [+referential] feature is checked in SpecDP: if the feature is "strong", the movement of the demonstrative is overt (as is the case for Italian), while if it is "weak", the movement can be procrastinated to LF (as happens in Spanish).

Thus, the present proposal argues for analyzing the specific marker *edin* in the same place in which demonstratives are found (or, at least, in which they are interpreted cross-linguistically). According to Brugé's proposal, this would follow naturally. Demonstratives check their [+referential] feature in SpecDP. Hence, one is led to assume that *edin*, being referential, checks its [+referential] feature in the same position. The difference between Spanish and Bulgarian lies in the strength of the referential feature to be checked. While in Spanish the feature is weak and allows the procrastination of the demonstrative movement at LF, in Bulgarian the [+referential] feature is strong and obligatorily forces the movement of the demonstrative at the left edge.

²³ A search on the Bulgarian National Corpus for the combination *edin njakakâv* / *njakakâv edin* leads to six total results, four of which have the two items separated by a comma or an ellipsis, signaling that *edin* and *njakakâv* belong to two different phrases. It is possible to think that they marginally co-occur in emphasized contexts with a numeral reading of *edin*, as happens with demonstratives (cf. §5.2.2).

Furthermore, the fact that these two items are found in the same position consequently leads to the impossibility of their co-occurrence. This is intuitively true, as the demonstrative introduces a definite NP, while *edin*, despite being referential, marks indefiniteness. The fact that they bear complementary pragmatic features (the referent is supposed to be known to both the speaker and the hearer in the case of the demonstrative, as opposed to *edin* which entails only the speaker's knowledge) is, however, not a reason to postulate different positions for these two items which are both anchored to the speaker. Evidence corroborating this claim comes from an apparent counterexample in which *edin* does cooccur with a demonstrative, as in (20a) (see p. 175) and in (38) below.

- (38) a. No тази *edna* minuta bez spomeni e strašna.
 but this one.F minute without memories is scary
 'But this (single) minute without memories is scary.'
 (bgTenTen12, 211437990)
- b. Za тази *edna* godina v Nju Jork sa izdadeni pone
 in this one.F year.F in New York are given at.least
 8200 svatebni licenza za brakove između xora ot
 8,200 wedding license for weddings between people from
 ednakâv pol.
 same sex
 'In this single year, at least 8,200 same-sex marriage licenses were
 issued in New York.'
 (bgTenTen12, 93014140)

The examples in (38) show that *edin* occurring with a demonstrative is not interpreted as an indefinite marker, but rather as a numeral providing the cardinality of 'one'. In this case, the demonstrative bears the [+referential] trait and checks it in SpecDP (letting it percolate down the structure). Since the left edge of the nominal expression is already occupied, specific *edin* cannot surface; instead, *edin* can appear lower in the structure in the projection hosting numerals.

Moreover, *edin* and demonstratives show a certain similarity from a morphosyntactic point of view. In Bulgarian, demonstratives do not generally co-occur with the definite article²⁴ and cannot be preceded by other nominal modifiers, as they are arguably the occupants of the higher edge of the nominal expression. The same holds for the specificity marker *edin*. The sentence

²⁴ As mentioned in fn. 11, Bulgarian displays phenomena of MD (Rudin 2019; Franks 2020). From the point of view of the theoretical framework adopted here, MD is accounted for without any further stipulation.

in (39) is taken from the questionnaire; it is safely taken to be grammatical (mean score: 5.87 (on a scale of 1 to 7, cf. §5.2 above)), even though it looks like a counterexample to what has just been stated.

- (39) Poznavam dvamata sinove na Elisaveta:
 know.1SG two.DEF sons of Elisaveta
edini-jat uči medicina, a drugi-jat—pravo.
 one-DEF studies medicine while other-DEF law
 'I know Elisaveta's two sons: the one studies medicine, and the other studies law.'

In (39) *edin* appears in the articulated long form (*edinijat*). In this case, however, it can be argued not to function as a specificity marker: the referent is not presented as indefinite, as the numeral picks an entity out of a set which has been previously specified and is therefore familiar to the hearer (*dvamata sinove* 'the two sons'). Moreover, the articulated form of 'one' always entails that the single entity is part of a larger (specified) group (Iliyana Krapova, p.c.). In the context of (39), 'one' is used to create a contrast with the second construal, as the opposition *edinijat* – *drugijat* (analogous to English *the one* – *the other*) clearly shows. On the other hand, the co-occurrence with a higher modifier, such as a possessive, shown in (40), is ungrammatical. Sentence (40) scored 1.83 on average in the questionnaire.

- (40) *Imam mnogo prijатели ot različni točki na sveta.
 have.1SG many friends from different points of world.DEF
 Moja *edna* brazilska prijatelka ne može da mi dojde
 my one.F Brazilian friend.F not could *da* I.DAT came
 na gosti minaloto ljato.²⁵
 on guest last.DEF summer
 Intended: 'I have many friends from all over the world. A Brazilian (female) friend of mine couldn't come to visit me last summer.'

²⁵ An anonymous reviewer points out that the presence of the definite article on the possessive would make the sentence grammatical (i.e., *mojata edna brazilska prijatelka*...). However, as Iliyana Krapova (p.c.) pointed out, this is marginally possible and only with the meaning 'my only Brazilian (female) friend'. Hence, in such a case, *edin* would play the role of the numeral and would be interpreted lower in the structure.

The fact that *edin* in this instance cannot be preceded by the possessive is easily explained by the fact that the former arguably occupies the highest nominal projection.²⁶

5.2.3. The Semantics of *edin*

The comparison that was drawn between indefinite referential *edin* and the demonstratives finds further support in a cross-linguistic perspective by looking at their semantics and the conditions that license their occurrence.

Bernstein (1997: 95) observes that demonstratives are ambiguous between a deictic and an “indefinite specific” reading, as in (41):

- (41) a. This woman (right here) (deictic)
 = this woman
- b. This woman (from Paris) (indefinite specific)
 = a woman

Ionin (2006) analyzes the conditions which license the occurrence of the “indefinite specific” *this*, building on Fodor and Sag’s (1982) notion of referentiality. Crucially, Ionin analyzes the specificity associated with the use of “indefinite” *this* as including a noteworthiness condition, which plays a crucial role in its licensing. This property hinges upon the speaker’s manifestation of a certain degree of knowledge about the referent, or upon the fact that the referent itself displays some noteworthy property. Ionin (2006: 185) provides the examples reported in (42–43).

- (42) a. #I want to see *this new movie*.
 b. I want to see *this new movie* that my friends have been
 recommending to me for ages.

²⁶ In the framework adopted here, possessive adjectives are also occupants of the SpecDP position, so one might wonder how *edin* and a possessive adjective may co-occur. Possessive adjectives, as already acknowledged by Giusti (2002: 144), are merged lower in the structure and are moved for interpretative reasons to SpecDP only if that position is not already occupied by another element. Giusti argues that the original position of the possessive is SpecNP, as nouns such as some kinship terms assign a θ -role in that position. Let us imagine a structure containing both *edin* and a possessive in their base positions. Let us also assume that some feature on the head D acts as probe and looks for a goal that may check the feature required for the interpretation of the whole expression in SpecDP. Given their base positions, the numeral is the first available goal that the probe meets, and it is attracted to SpecDP. The possessive will be able to move but lands in the specifier of a lower reprojection of N, as happens in Italian when the possessive is preceded by the article (cf. Giusti 2015: 151 for a representation of this movement).

- (43) a. I found *this blue apple* on my plate!
 b. #I found *this apple* on my plate!

In (42b) the speaker's knowledge about some facts related to the movie (expressed by the relative clause modifying *this new movie*) suffices to license the use of non-referential *this*, contrary to what happens in (42a). What licenses the use of indefinite *this* in (43a), but not in (43b), is the adjective *blue*, which defines the noteworthy property of the apple.

The same use of "indefinite" 'this' is possible in Bulgarian as well. In a context in which there is no film to be deictically pointed at (e.g., a situation in which two friends are talking while walking in a park), (44) is infelicitous if uttered out of the blue and without any further information about the film in question. Sentence (45) is instead perfectly grammatical, as the key information licensing the use of the "indefinite" demonstrative is present.²⁷

- (44) #Iskam da gledam *tozi* film.
 want.1SG *da* see.1SG *this* film
 'I want to see *this* film.'
- (45) Namerix *tazi* kniga, za kojato mi govoreše.
 found.1SG *this* book for which I.DAT talked.2SG
 'I found *this* book you told me about.'

At first glance, it seems that in Bulgarian the conditions licensing the occurrence of specific *edin* correspond to those posited for English *this*-indefinites. Let us take a look at sentences (46–49), which exemplify the licensing conditions for *edin*.

- (46) Sâprugata na Ivan ima kovid. Toj e pritesnen i iska
 wife.DEF of Ivan has Covid he is worried and wants
 da govori s *edin* lekar, d-r Borisov, može bi
da talks with one.M doctor Dr. Borisov can COND
 go poznavas?
 him know.2SG
 'Ivan's wife has Covid. He is worried and wants to talk to a doctor, Dr. Borisov, maybe you know him?'

²⁷ I thank Assia Assenova for having provided me with the examples in (44) and (45).

- (47) Čux, če *edin* lekar otriča săștestvuvaneto na kovid.
 heard.1SG that one.M doctor denied existence.DEF of Covid
 Iskam da znam koj e toj.
 want.1SG da know.1SG who is he

'I heard that a doctor denied the existence of Covid. I want to know who he is.'

- (48) Elena pročete vsički knigi, koito ì preporâča
 Elena read all books which she.DAT recommended
edin prepodavatel. Ne znam koj e toj.
 one.M professor not know.1SG who is he

'Elena read all the books that a professor recommended to her. I don't know who he is.'

- (49) *Sâprugata na Ivan ima kovid. Toj e pritesnen i iska
 wife.DEF of Ivan has Covid he is worried and wants
 da govori s *edin* lekar, kojto i da e toj.
 da talks with one.M doctor who and da is he

Intended: 'Ivan's wife has Covid. He is worried and wants to talk to a doctor, whoever he is.'

In (46) *edin* is referential in the most traditional sense of the term: the speaker is able to identify the referent, whose name is mentioned in the discourse. This sentence was accepted with a mean score of 5.41. As for the remaining examples, the noteworthiness requirement is satisfied in (47) and (48), but not in (49). In the latter, the speaker does not exhibit any knowledge about the doctor she is talking about, whose identity cannot be established, as shown by the expression *kojto i da e toj* 'whoever he is'. This sentence scored on average 2.76. In (47), too, the referent is not directly identified by the speaker, as revealed by the second statement (*Iskam da znam koj e toj* 'I want to know who he is'). In this case, however, there is a noteworthy property about the person who is being talked about, namely, the fact that he denied the existence of Covid (which is quite surprising given the fact that this person is a doctor). This noteworthy property licenses the occurrence of *edin*, as shown by the mean judgment score of 5.70. Similarly, (48) is accepted with a mean score of 5.65. Even if the action of recommending a book is not noteworthy in the traditional sense, it shows that the speaker has the knowledge of some property associated with the referent, in contrast with the ungrammatical (49). This is the property which specific *edin* shares with the "presentative marker" *edin*, which typically occurs at the beginning of fairy tales (see (5) on p. 166). In these contexts, in

fact, the narrator displays a certain amount of knowledge about the referent introduced by *edin*.

The current subsection has drawn a parallel between “indefinite” *this* (and *tozi*) and specific *edin*, showing that they pattern alike with respect to their licensing conditions. Their semantic similarity is another element which encourages a unified treatment of these items at the level of syntax as well. This is further empirically justified by a comment by Assia Assenova (p.c.), who observes that *edin* in the sentences (46–48), but crucially not (49), can be substituted by “indefinite” *tozi* ‘this’. Not only can these the two items be substituted for each other, but they are also in complementary distribution, as shown in (50).²⁸ This fact strongly suggests that they occupy the same position inside the nominal expression.

- (50) a. Pročeto*x* *tazi* kniga, za kojato ti govoreše.
 read.1SG this book about which you.NOM talked.2SG
 ‘I read this book you were talking about.’
- b. Pročeto*x* *edna* kniga, za kojato ti govoreše.
 read.1SG one.F book about which you.NOM talked.2SG
 ‘I read a book you were talking about.’
- c. Pročeto*x* *tazi* (#*edna*) kniga, za kojato ti govoreše.
 read.1SG this one.F book about which you.NOM talked.2SG
 ‘I read this (particular) book you were talking about.’

This section discussed the use of *edin* as a specificity marker, arguing that its position is likely to be the same as that of demonstratives, i.e., SpecDP, the left edge of the nominal expression. The next section will take into account the last stage of the development of *edin*, in which it behaves as an article-like item that is interpreted non-referentially.²⁹

6. *Edin* as an Article-Like Element

As argued above, the numeral *edin* partially shows properties of the stage of the indefinite article. Clearly, as was stressed by Geist (2013), *edin* cannot

²⁸ As an anonymous reviewer points out, it is possible to have the demonstrative and *edin* co-occurring. In such a case, however, *edin* receives the numeral interpretation that was mentioned in example (20).

²⁹ Here I refrain from labeling this occurrence of *edin* an “indefinite article”, as the grammaticalization path predicts. Thus, even though this development is expected, I will not commit myself in attributing to it the status of an indefinite article. What is relevant for the present discussion is its sharing of some features with articles.

be said to be a fully-fledged indefinite article at this stage. However, in this section, I will argue that *edin*, in its article-like function, already occupies the position traditionally assigned to articles, i.e., the head of the DP. Section 6.1 traces a parallel between non-specific *edin* and definite articles. Section 6.2 underlines the article-like properties of this use of *edin*, and §6.3 outlines the advantages of this theoretical approach.

6.1. Parallel with the Definite Article

Out of the three features displayed by a full-fledged indefinite article identified by Geist (2013), Bulgarian *edin* only displays one, i.e., the ability to appear in a generic context with non-referential interpretation. Let us take a couple of examples from the questionnaire: (51) was accepted with a rate of 5.61, while (52) had an average score around 5.35.

- (51) *Edna žena vinagi e prava.*
 one.F woman always is right
 ‘A woman is always right.’

- (52) *Statistički edin bâlgarin živee sredno 75 godini.*
 statistically one.M Bulgarian lives on.average 75 years
 ‘Statistically, a Bulgarian lives on average 75 years.’

In these examples, *edin* is interpreted in a completely different way with respect to the interpretation it receives when it is a specificity marker. The sentence in (51) is commonly interpreted as equal to ‘women are generally always right’. Similarly, (52) conveys the meaning that ‘Bulgarians generally live on average 75 years’. In this latter case, the referential reading which could have arisen in the former sentence³⁰ is prevented by using the adverb *statistički* ‘statistically’. This amounts to saying that, in these instances, *edin* refers to a prototypical representative of the class of referents denoted by the NP (women and Bulgarians, respectively) rather than picking a specific entity out of the denotation of the NP. This use of *edin* is not limited to sentence-initial position, as suggested by (53) (taken from (11) above), which was judged perfectly grammatical by the native speakers consulted.

³⁰ Assia Assenova (p.c.), however, assures me that the most natural reading of sentence (50) is one in which *edin* is non-referential. A referential reading could of course arise, but only if the context contains a clear indication that *edin* is referring to a woman in particular.

- (53) *Edin* džentâlmen vinagi otvarja vrata na *edna* dama.
 one.M gentleman always opens doors to one.F lady
 'A gentleman always opens doors for ladies.'

The occurrence of non-specific *edin* in generic sentences is constrained to those sentences with a predicate selecting a non-kind-referring NP: thus, *edin*-NPs cannot appear as subjects of predicates such as *be extinct*, as is the case with English NPs introduced by the indefinite article (Krifka et al. 1995).

The fact that *edin* in these instances does not refer to a particular referent already indicates that it cannot occur in the same position as referential *edin*, as it would be unlikely that the same element could receive two different interpretations in the same position. More so, sentences like (51) are potentially ambiguous between a referential and non-referential interpretation; thus, finding the source of this difference in the syntax would be a welcome result.

I argue that this difference in the interpretation of *edin* is due to its diachronic reanalysis from SpecDP to D. As seen in §3.2, the D position in Bulgarian is specified with the [DEF] trait, with which the definite article agrees in a local configuration to be interpreted. It is interesting, but not surprising, to notice that the same sentences in (51–52) may be rephrased with a singular or plural NP inflected for the definite article (shown in (54a–55a) and (54b–55b), respectively). The resulting sentences obtain the same interpretation as those featuring *edin* (if a semantic difference is there, it is very subtle),³¹ as confirmed by the informants. Sentences featuring specific *edin* (e.g., (47)) cannot be rephrased using the definite article, since the referent would instead be presented as known to the hearer, as reported by the consulted informants for (56).

- (54) a. Žena-ta vinagi e prava.
 woman-DEF always is right
 b. Ženi-te vinagi sa pravi.
 women-DEF always are right
 'Women are always right.'

³¹ It is known that different kinds of NP display different distributions with respect to the predicate they can combine with (Krifka et al. 1995). However, an extensive discussion on this issue would lead us too far away from the current topic. For the purpose of the present work, it is sufficient to know that *edin*-NPs can be interpreted non-specifically in a subset of generic sentences and that native speakers tend to interpret these in the same way as they would interpret NPs inflected for the definite article. A complete discussion of the subtle differences among different kinds of NP appearing in generic sentences in Bulgarian is left for future research.

- (55) a. Statističeski bǎlgarin-ât živee sredno 75 godini.
 statistically Bulgarians-DEF lives on.average 75 years
 b. Statističeski bǎlgari-te živejat sredno 75 godini.
 statistically Bulgarian-DEF live on.average 75 years
 ‘Statistically, Bulgarians live on average 75 years.’
- (56) *Čux, če lekar-jat otriča sâštestvuvaneto na kovid.
 heard.1SG that doctor-DEF denied existence.DEF of Covid
 Iskam da znam koj e toj.
 want.1SG da know.1SG who is he
 Intended: ‘I heard that the doctor denied the existence of Covid. I want to know who this is.’

This observation that the semantic contribution of generic *edin* is in this case (almost) equivalent to that of the definite article further strengthens the assumption that generic *edin* occurs in the head D, which is associated with (and also diachronically related to) the position of the definite article (cf. Giusti 1995).

This account could *prima facie* run into a problem: if generic *edin* is realized as the head D, the specifier remains empty and could then host a demonstrative. Still, the demonstrative cannot co-occur with generic *edin*, as it would force a specific reading of the indefinite, thus picking a single entity out of the set denoted by the NP, ultimately losing the non-referential interpretation. The impossibility of co-occurrence of the two elements is, however, naturally ruled out by the Doubly Filled XP Filter: the functional projection DP needs to be visible by realizing one of the two positions associated with it. Once *edin* is realized, the head D is overtly spelled out, and the specifier position need not (for Economy reasons, cannot) be overtly realized.

Another parallel with definite articles that supports the idea that non-specific *edin* is the product of the reanalysis of the specifier of the DP into the head D is the similarity of this process with that which led to the development of the definite article in Romance.³² In particular, Giusti (2001) argues that the Italian definite article developed from the morphological weakening of the Latin demonstrative *ille* in SpecDP. Its reduced form was reanalyzed as the head D, as shown in (57), which is the bracket notation of Giusti’s (2001: 167) representation.

³² The shift from the demonstrative to the article holds in this case as well: the Bulgarian definite article is diachronically derived from the old Slavic demonstrative pronouns in unstressed position, which cliticized onto the noun (cf. Mangiulea 1987). I thank the anonymous reviewer for pointing this out.

- (57) [_{DP} IL(LE) [_D Ø]] > [_{DP} Ø [_D IL(LE)]]

This gives strength to the proposed analysis, showing that the reanalysis of SpecDP into the head D is not an isolated phenomenon and that this same mechanism is responsible for the development of a grammatically related item in a quite unrelated language. Another interesting detail, which further supports the view of generic *edin* in D, is to be mentioned. As Giusti (2001: 197) admits, given the absence of lexical material in either the specifier or the head of the DP, the two structures may have coexisted for several generations. This is the case for specific and nonspecific *edin*, whose positions are hardly distinguishable if we look at the superficial level.

6.2. Article-Like Properties of Non-Specific *edin*

Superficially, non-referential *edin* occurring in generic sentences also displays two important features that are linked to articles: its phonetic weakness and its syntactic dependency on the head N (Giusti 1997).

As far as the former property is concerned, Alexander (2000: 55) (quoted in Leafgren 2011: 61) states that “[w]hen *един, една, едно* [*edin, edna, edno*] means ‘a’, Bulgarians tend to pronounce it with a much weaker accent than when it means ‘one’”. This indicates that *edin* is undergoing a process of phonological weakening, which is expected to ultimately lead to morphological erosion, as happened in the Slovene dialects of Friuli (Benacchio 2018).³³ Moreover, another piece of evidence in favor of the weakened phonetic form of non-specific *edin* is the impossibility of focusing it without producing a change in its interpretation. If *edin* is focused in a sentence like (58), it gets a specific reading (Iliyana Krapova, p.c.).

- (58) *EDNA* *žena* *vinagi* *e* *prava*.
 one.F woman always is right
 ‘A certain woman is always right.’

As for the syntactic dependence from the head N, it was shown that specific *edin* could license a null nominal (cf. (37b)). In the instance under investigation, however, it is not possible to separate non-specific *edin* from the N it introduces. Sentences like those in (59) are not interpretable in a non-specific

³³ In the Slovene dialects spoken in Friuli, “together with the accented forms of the numeral (*dyn, dnö, dnä; dny, dne*), we also have the corresponding clitic forms, in proclitic position, before the noun phrase. These forms mostly lack the initial phonetic element -d (*din/ni, nö/nu, na; ni/ne*). These forms no longer have a quantitative function, but confer an indeterminate value to the noun phrase, evidence that the referent is unknown to the listener” (Benacchio 2018: 205).

way, as the informants pointed out. Similarly, articles cannot be separated from the N they co-occur with (cf. (38)).

- (59) a. [Talking about women]
 **Edna* vinagi e prava.
 one.F always is right
 Intended: 'A woman is always right.'
- b. [Talking about Bulgarians]
 **Statističeski edin* živee sredno 75 godini.
 statistically one.M lives on.average 75 years
 Intended: 'Statistically, a Bulgarian lives on average 75 years.'

These are further pieces of evidence pointing in the direction of analyzing non-specific *edin* as a functional head realizing D, parallel to other articles. Let us now turn to some advantages of this analysis.

6.3. Advantages of Non-Specific *edin* in D

As far as the syntax-semantics interface is concerned, a strong reason to suppose that non-specific *edin* is in the head D is that it does not contribute to the interpretation of the noun it occurs with (as it does not pick any referent out of the denotation of the NP, whose extension remains unaffected). In this sense, its interpretation in the generic sentences it appears in is almost analogous to the interpretation of the definite article. Moreover, taking non-specific *edin* to be the spell-out of the head D has the advantage of providing an account for its co-variance with the structures displaying the definite article. Recall sentences (51) and (54a), repeated here in (60).

- (60) a. *Edna* žena vinagi e prava.
 one.F woman always is right
 'A woman is always right.'
- b. *Žena-ta* vinagi e prava.
 woman-DEF always is right
 'Women are always right.'

As shown by D-V&G, Bulgarian Ns never undergo N-to-D movement in overt syntax, but they do in LF if they are "bare", i.e., if there are no other nominal modifiers and thus the definite article is found on the N itself, as is the case in (60b). The movement is necessary to make the DP visible at the interface, at

the same time checking the features of the definite article. In the case in which *edin* is inserted in D, no movement of N is possible in LF, as the position is already occupied. The movement is, however, not necessary, as the functional category DP is already visible at the interface. The two LF structures are given in (61).

- (61) a. [DP [D žena-ta] [NP ~~žena~~-ta]]
 b. [DP [D edna] [NP žena]]

Assuming (61), the two structures are likely to coexist in that they require the same derivational cost (cf. Biberauer and Richards 2006), and the system is not able to decide which one is more economical. In (61a), N features the definite article and needs to move in LF to the DP in order for the article to be interpreted by making the projection visible at the interface with semantics. This movement is not necessary in (61b), in which *edin* is directly merged in D, making the movement of N unnecessary. Note that the two constructions do not imply any violation of the Merge-over-Move Principle (cf. Chomsky 1995): in fact, the movement involved in (61a) is more economical in that it happens at LF, according to Procrastinate.

Assuming non-specific *edin* in D also lends support to the hypothesis that specific *edin* is in SpecDP. As was already pointed out, specific *edin* is assumed to behave like a demonstrative, being endowed with a [+referential] feature which is checked in SpecDP. This feature manifests itself in the possibility of the speaker (or of the bearer of attitude) to individuate the referent (or some noteworthy property of it) introduced by *edin*. Giusti's (2002) claim that the interpretation of the nominal expression takes place in SpecDP at LF translates in the assumption that the [+referential] feature can be checked only at the very left edge of the nominal expression. If an element does not appear in that position, it will not be able to have the [+referential] feature checked. Assuming this, the position of non-specific *edin* in D is naturally accounted for: as it does not appear in SpecDP, it does not get a specific interpretation.

Another theoretical advantage of differentiating the two positions inside the DP (specific *edin* in SpecDP and non-specific *edin* in D) is that it is possible to trace a straightforward parallel between the nominal left periphery and the clausal one (a parallelism already drawn by Abney 1987 and Longobardi 1994, inter alia). Interestingly enough, the assumption that the elements anchored to the speaker sit in SpecDP finds an interesting parallel with Giorgi's (2009, 2012) theory on the representation of the speaker's coordinates in the left periphery of the clause. Giorgi (2009) argues that the speaker's space-temporal coordinates are represented in syntax in the form of a deictic element pointing at the speaker. In a split-CP model (cf. Rizzi 1997), this deictic element is

realized in the highest layer, above ForceP.³⁴ Thus, the highest projection of the left periphery of the clause is the locus where the tense of the utterance is “anchored” to the speaker. Following the proposal assumed here, the left edge of the nominal expression would carry out the same function as the left edge of the clausal left periphery. As nominal expressions, unlike clauses, lack Tense (cf. Giusti 2006), the features that can be checked in SpecDP are arguably those which have to do with the speaker in the nominal domain, i.e., referentiality (and spatial deixis).

One possible issue raised by the present proposal is the morphophonological equivalence between *edin* sitting in SpecDP and *edin* sitting in D, implicating, instead, a difference between their properties, as the former is a specifier, while the latter is a head. Mainly, what remains to be accounted for is, on the one hand, the fact that *edin* in the head D still agrees with the nouns it modifies (while this is usually a trait of phrases sitting in specifier position which undergo Concord), and on the other hand, the fact that there is no morphologic erosion of this latter element. As for the latter, §6.2 showed that *edin* is undergoing a process of phonetic weakening. Moreover, as was pointed out, *edin* has not reached the status of a full-fledged indefinite article. Hence, one is led to conclude that the reduction of *edin* will occur in a more advanced stage of development of the article.

As far as the agreement issue is concerned, the situation is less dramatic if looked at from the perspective of the structure of nominal expressions. The assumption, presented in §3.1, is that the nominal spine is created by re-emerging the functional features of the lexical noun; this creates functional heads containing a copy of the features of the nominal expression. Since *edin* overtly

³⁴ Giorgi (2012: fn. 3) tentatively labels this projection “C-Speaker”, which is realized by the complementizer *che* ‘that’ introducing an indicative subordinate in Italian. In such case, shown in (i), it is not possible to delete the complementizer, while it is optionally realized when it introduces a subjunctive subordinate, as in (ii).

(i) Gianni ha detto *(che) è incinta.
Gianni has said that is.IND pregnant
‘Gianni said that she is pregnant.’

(ii) Gianni credeva (che) fosse incinta
Gianni believed that was.SUBJ pregnant
‘Giannia believed she was pregnant.’

(Giorgi 2012: 45)

Interestingly, (i) triggers Double Access Reading (DAR), i.e., the subordinate tense is checked both against the main subject’s temporal coordinate (*Gianni*) and that of the utterer (i.e., *now*). This means that (i) is true if the state of pregnancy of the third person holds both at the time Gianni uttered the sentence and *now*. Example (ii) does not trigger DAR; consequently, the state of pregnancy is understood to hold only at the time Gianni uttered the sentence. Thus, *che* in (i) encodes the speaker’s temporal coordinate: it triggers DAR and cannot be deleted.

displays the φ -features of the head N, it is a perfect candidate for being reanalyzed as a mere spell-out of the nominal features. Since each functional projection needs to be made visible, I argue that non-specific *edin* is reanalyzed as a reprojection of the functional features of the noun, thus sitting in head position.

The present subsection has extensively argued in favor of considering non-specific *edin* to be the spell-out of nominal features in head D. Once all the pieces of the puzzle have been set down, we can take a step back to look at the general picture obtained from the present analysis.

7. Summing Up the Chunks

The analysis developed so far has been independently motivated. However, looking at the more general picture, it also presents a great advantage from a theoretical point of view. In fact, the proposed grammaticalization path of *edin*, as described here, represents a perfect linguistic cycle. Section 7.1 outlines the features of grammaticalization processes. Section 7.2 describes the general process of grammaticalization of *edin*, showing its similarity with other cycles. Section 7.3 shows that the grammaticalization of *edin* can be considered a full-fledged grammaticalization process.

7.1. What is Grammaticalization?

Grammaticalization is a diachronic process whereby an item α is reanalyzed from being (semi-)lexical to being (semi-)functional. One of the most well-known examples of grammaticalization is represented by the so called “Jespersen cycle” (Jespersen 1917), which describes the way in which negation develops in English and other Indo-European languages (e.g., French). The cyclical nature of this development is given by the fact that, when a lexical item has fully grammaticalized (thus losing its original function), a new lexical item is added to compensate for the loss of the original meaning. This new item may in turn undergo the same process, creating a cycle. For example, the French preverbal negator *non* is at some point reduced to *ne* and thus needs to be reinforced. The postverbal marker *pas* is optionally added to reinforce *ne*, but at some stage, *pas* is interpreted as the “real” negator and is obligatorily inserted. *Ne* becomes optional (as it does not bare negative features anymore) until it is dropped, and *pas* becomes the only negative marker. This process is often accompanied by a loss of phonological weight and semantic specificity (van Gelderen 2008). The cycle of negation in English (cf. van Gelderen 2013 and references therein) follows a similar path.

It is of great interest to investigate the syntax that underlies these changes and the principles that guide the cycle, as they can perfectly account for the

change of Bulgarian *edin* from numeral to specificity marker, and ultimately to article-like marker of genericity (see §7.2).

Collecting together what was proposed in her earlier works, van Gelderen (2011) proposes that linguistic cycles are driven by the reanalysis of formal features (introduced by Chomsky 1995), which are accessible during the derivation. They may be either interpretable or uninterpretable. The former ones are readable by the semantic interface, while the latter are not and thus need to be checked and eliminated. In linguistic cycles, interpretable features are reanalyzed as uninterpretable ones, which are more economical in that they cause the derivation to proceed in order for them to be eliminated. The driving principle is stated by van Gelderen (2013: 246) as in (62), in which she provides the example of the negation cycle in English.

(62) Feature Economy

Minimize the semantic and interpretable features in the derivation:

DP in the VP		Specifier of Neg		Head Neg		Negative affix
Semantic	>	[iF]	>	[uF]	>	[uF] ³⁵

The negation cycle in English can be captured by the stages described in (62): the typical Old English negator *ne* ‘not’ bears negative interpretable feature [iF], which is subsequently analyzed as [uF], and thus acts as a probe searching for a goal to attract in the specifier of the Neg(ation)P(hrase) to eliminate the [uF]. In Middle English, the negative argument *nowuth/nan wuht* ‘no thing’ is promoted to SpecNegP, as it bears a negative semantic feature. Afterwards, the negative [iF] of the specifier is reanalyzed as [uF], and it reduces and shifts to the head Neg position. In this way, the cycle is fed and goes on in the same way.

In cycles, mainly two principles come into play, reported in (63) and (64) (from van Gelderen 2011: 13–14).

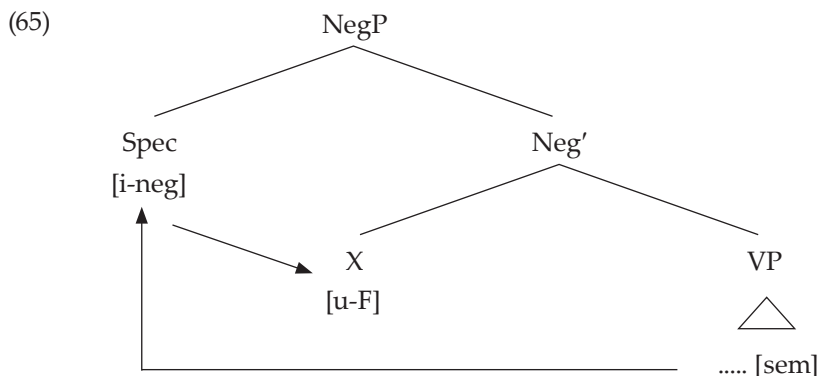
(63) Head Preference Principle (HPP): Be a head, rather than a phrase.

(64) Late Merge Principle (LMP): Merge as late as possible.³⁶

³⁵ “Semantic” refers to “semantic features” of the predicate, e.g., the verb *remain* has the semantic feature [duration] (van Gelderen 2013: 242). “iF” stands for “interpretable feature”, while “uF” indicates “uninterpretable feature”.

³⁶ For the sake of consistency, I will refer to the two Economy principles reported here. As pointed out by an anonymous reviewer, however, these principles are not entirely new in the literature on syntax: HPP is equivalent to Cardinaletti and Starke’s (1999) “Minimize Structure”, while LMP is grounded on the same principle underlying Chomsky’s (1995) “Procrastinate”. It should be pointed out, however, that LMP is not equivalent to Procrastinate, which favors LF movement over movement in overt

As far as the cycle of negation is concerned, LMP is responsible for the merging of the negative adverb from its base position to a higher one, i.e., SpecNegP. This complies with general Economy, as directly merging a phrase in a high position is less costly than merging it in a lower position and remerging the element higher in the structure. Once the negative adverb is merged in SpecNegP, the HPP intervenes in reanalyzing the specifier as a head. The whole negative cycle is graphically represented in (65) (reported here from van Gelderen 2013: 246).



The model provided in (65) is consistent with a formal approach to grammaticalization (cf. Roberts and Roussou 2003), in that it involves a structural shift “upwards” in the functional hierarchy, which then involves a loss of movement (as the element is directly merged higher in the structure). Moreover, this process goes hand in hand with phonological reduction (from *nowuth/nan wuht* to *not*) and semantic bleaching of the original element (which, in van Gelderen’s terms, loses its “semantic features”).

One last thing to be mentioned to conclude this section is that there are three core features which distinguish grammaticalization (Diewald 2011: 367): (i) paradigmatic integration, (ii) obligatoriness, and (iii) relational meaning. The property in (i) refers to the Jakobsonian idea that a grammaticalized item is, by definition, the marked member of an opposition with a notionally unmarked zero element and, as such, becomes a member of a paradigm (cf. Diewald 2011). In virtue of that, (ii) expresses the necessity of operating a choice between the opposite values of the paradigm, i.e., the information conveyed by this opposition needs to be expressed by choosing one of the two items (either the grammaticalized one or the notional zero). Importantly, the

syntax. LMP is instead intended by van Gelderen as a principle favoring the direct merge of a phrase in a higher structural position rather than its low merge followed by its upward movement.

obligatoriness requirement is a matter of degree: the grammaticalized element may be obligatory in some contexts but not in others (Diewald 2011; Lehmann 1995/1982). The property in (iii) assumes that grammatical signs create a link between the element they modify and another entity, which typically results in an indexical relation (e.g., with the speaker). A typical example of (iii) is subjectification (cf. Traugott 1989, 1995), in which a grammaticalized item comes to encode the speaker's perspective (e.g., the hortative *let's* developed from the imperative construction *let us*).

7.2. What About *edin*?

As pointed out in §1.2, the development of different functions of *edin* is qualitatively different from the process of grammaticalization depicted by the Jespersen cycle.³⁷ In the case of negation, the newly grammaticalized negative element supersedes the old one, thus maintaining a sort of “equilibrium” inside the lexicon: one item is gained, but another one is lost. Things are different when looking at *edin*: the newly grammaticalized item(s) do(es) not cause the loss of the original one, i.e., the numeral. However, there are many reasons to consider this as a process of grammaticalization with full rights.

Assembling the pieces of the proposal outlined in the previous sections, one notices that the syntactic analysis proposed for the numeral *edin* in Bulgarian is perfectly in line with this model of grammaticalization of the negator in (62). The path à la van Gelderen is summarized in (66).

(66) The cycle of Bulgarian *edin*

SpecNumP		SpecDP		Head D
[iF]	>	[iF]	>	[uF]

In this view, *edin* starts out as a numeral, and in its grammaticalization process, some of its features are reanalyzed and some others are lost. It is important to point out that the model I am proposing here is to be taken as a diachronic one, since grammaticalization is a process which unfolds over time.³⁸

³⁷ I am referring here to *edin*, but the same reasoning straightforwardly applies to the process of grammaticalization of the numeral ‘one’ cross-linguistically.

³⁸ I am not claiming a sort of “online grammaticalization”, i.e., that *edin* always enters the numeration as a numeral and then undergoes movement based on its function. What I am trying to do here is look at synchrony to understand which syntactic changes may have occurred in the linguistic development of Bulgarian. This provides a model which can be used to guide a corpus study in search of diachronic data supporting or refuting the hypothesized development of *edin*.

A question, however, arises: why is *edin* eligible to undergo this process? The answer is already provided by Givón (1981: 51):³⁹

Quantifying expressions [...] *imply referentiality but do not imply prior-acquaintance/familiarity*. They are thus the only major class of noun-modifiers in the NP that fulfils the requirement for the development of a referential-indefinite marker. (*italics mine*)

Following Givón's suggestion, the numeral *edin*, in its cardinal function, is endowed with an interpretable quantificational feature [+quant] roughly indicating that the set of the referent contains only one element. This feature also implies referentiality (indicated as [+ref] in the representation in (67)), as the entity needs to have a reference to be quantified.

Let us suppose that in specific contexts nominal expressions with a singular head N have a referential [uF] in D which acts as a probe,⁴⁰ searching down the tree for the first suitable goal bearing a referential feature [+ref], finding it in the numeral *edin*. As a consequence, *edin* is attracted to SpecDP, checking and eliminating the [uF] on the head. This probe-goal relation (and consequent movement of *edin* to SpecDP) is repeated over time. Once this movement is well established, LMP applies, merging the item directly in SpecDP, which is more economical than remerging the item from a lower to a higher position. This is an instance of what has thus far been called "specific *edin*". After this stage, some specific syntactic configurations (i.e., subject position of generic sentences with individual-level predicates) create favorable conditions for the application of HPP, which causes the interpretable referential feature of *edin* in the specifier to be reanalyzed as an [uF] in the head D. This follows naturally from Giusti's assumption that articles are bundled with (abstract) Case features (cf. §3.1), and Case is per se uninterpretable (cf. Giusti 2011). This [uCase] feature is checked by the head selecting the DP, e.g., a lexical verb. As said before, this stage of the process (the reanalysis from specifier to head) is not generalized to all occurrences of *edin*, which is a sign that this item has only quite recently entered this stage. This is probably the reason why morphological erosion has not applied yet, although at the present stage, it is

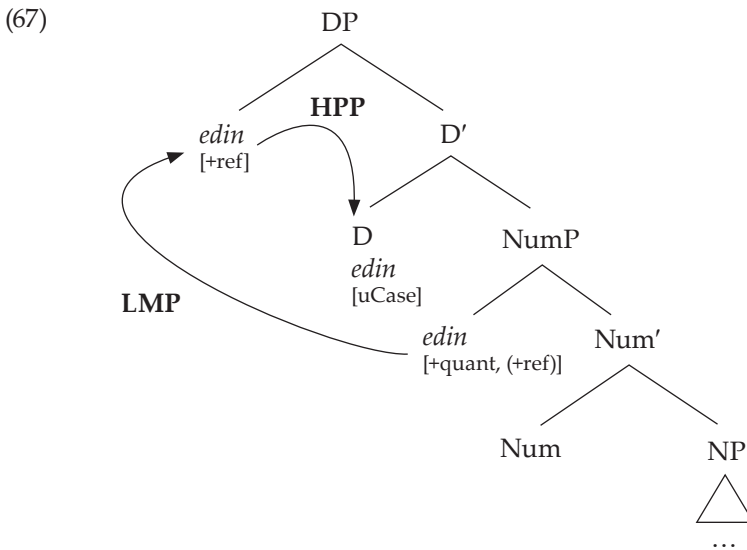
³⁹ An interesting proposal by Crisma (2015) regarding the grammaticalization of *a(n)* in English is that the numeral 'one' is the only cardinal having both the lower bound ('at least one') and the upper bound ('at most one') as part of the lexical meaning (while in the other numerals, the upper bound is only implicated; cf. Crisma 2015 and Horn 1972). As such, 'one' has a special status, and both the upper and lower bound may be bleached during the process of grammaticalization. I will not dwell on this inspiring proposal here, and I direct this question to future research.

⁴⁰ This is in line with Longobardi's (1994) proposal that the DP layer is independently needed for referential requirements.

possible to trace a phonetic reduction, with *edin* in generic sentences losing the independent accent and being unable to be focused.

If there is still no (superficial) reduction of *edin*, the process of semantic bleaching is instead attested. Givón's (1981) insight is particularly interesting in this respect: the feature distinguishing numerals is that they not only express cardinality but also imply referentiality. The process of semantic bleaching of the numeral 'one' seems to involve the gradual loss of the quantitative semantics to leave space for the referential one. This could be described by the Langackerian metaphor of glasses (although not in the sense of Langacker 1990). If we wear glasses and we focus on some external object, the construal of the glasses fades away from our consciousness and we no longer notice it. Something similar is likely to happen with 'one': the referential implication becomes the "object of attention", causing the rest (in this case, the quantitative semantics) to fade away.

Coming back to *edin*, it is possible to notice that there is a striking analogy in the syntactic positions between linguistic cycles and the "path" of grammaticalization which *edin* undergoes. The latter is exemplified in (67), which is analogous to the representation given by van Gelderen (2013) of the negation cycle reported in (65).



This analogy with the syntactic model of grammaticalization of negation in (63) seems a good reason to believe that the analysis sketched up to this point may at least be on the right track. Moreover, this model is perfectly in line with Crisma's (2015) proposal of grammaticalization of the cardinal *an* in Old

English into the indefinite article *a(n)* in present-day English. Crisma (2015: 142) distinguishes three stages in the historical development of *an*:

- i. In Old English, it is a cardinal “merged in a quantity projection lower than D, say NumberP or #P”.
- ii. At a later stage, *an* becomes an existential operator marking specificity “found in (or in some position in the D-field)”.
- iii. In present-day English, *a(n)* as an indefinite article is an expletive merged directly in D.

The model presented here perfectly integrates with Crisma’s analysis of the grammaticalization of the indefinite article in English. This suggests that the model in (67) could in principle be extended to other Slavic and non-Slavic languages (e.g., those mentioned in §1.2, but also Romance languages) in which the numeral ‘one’ is grammaticalizing (or has grammaticalized) into an indefinite article.

7.3. An Instance of Grammaticalization

Before concluding, let us have a look at the general properties of the process to see which implications it has as far as the concept of grammaticalization is concerned. As shown in the previous sections, this process qualifies with full rights as an instance of grammaticalization. However, it must be acknowledged that it is qualitatively different from other processes which fall under the same umbrella term (e.g., the negation cycle). The difference lies in the fact that the new grammaticalized functions stemming from the numeral *edin* pile up and are added in the lexicon, without causing the loss of the source numeral.⁴¹

Leaving this difference aside for the moment, the development of *edin* displays the features that characterize grammaticalization. In Roberts and Roussou’s (2003) formal approach, the change *edin* undergoes involves an

⁴¹ I will not take any stand here about the mental representation of the newly grammaticalized functions of *edin*, i.e., whether they are represented as separate lexical items or whether it is a matter of underspecification of the only representation of *edin*. The case of English (and many other languages) would suggest the creation of a separate lexical item for the most grammaticalized functions (numeral *one* vs. indefinite article *a(n)*). This view poses a theoretical issue, i.e., adding new items to the lexicon without dispensing with the “old” ones is anti-economical. However, this process seems to be quite productive (an example is the creation of neologisms for new referents, e.g., *computer* or *to google*).

upward movement in the functional hierarchy, accompanied by semantic bleaching and a prelude of phonological weakening.

The process *edin* undergoes also fits perfectly within Diewald's (2011) characterization of grammaticalization (see §7.1). The new grammaticalized functions respect the three main features outlined above:

- i. Paradigmatic integration: *edin* as a marker of specific indefiniteness enters a paradigm in opposition to the zero marking (bare NPs). The two values of this paradigm—*edin*-NPs vs. bare NPs—is likely to encode the “token” vs. “type” opposition (cf. Gorishneva 2013).
- ii. Obligatoriness: once *edin* is grammaticalized as an indefinite marker or as an article-like element, its insertion is obligatory at least in some contexts, e.g., when introducing indefinite topics (cf. (10)) or in some cases in generic contexts (cf. (11), although in co-variance with the definite article).
- iii. Relational meaning: *edin*, from a marker of cardinality, grammaticalizes an indexical relation with the speaker, i.e., (non-)specific indefiniteness. In this sense, we could trace a process of *subjectification* alongside the grammaticalization of *edin*: from cardinal numeral, it becomes a marker of the attitude of the speaker towards the hearer. In fact, by definition, (non-)specific indefinites introduce referents (un)familiar to the speaker (familiar if specific, unfamiliar if non-specific) but which are presented as unknown/unfamiliar to the hearer. In this sense, the use of an indefinite reveals the speaker's inference that the hearer does not possess the knowledge to identify the referent that is being introduced in the discourse.

Acknowledging that the grammaticalization of *edin* also involves a certain degree of subjectification allows us to trace yet another interesting parallel: also in the case of *edin*, the new “subjectified” functions coexist with the original element, as is the case with items undergoing subjectification described by Traugott (1995).⁴² After all, cases of layering (i.e., the coexistence of both the grammaticalized item and the original source) are common in the numeral domain as well; as von Mengden (2008) argues, expressions for body parts are the main source for cardinal numerals. In some languages, the original

⁴² For example, the andative construction *to be going to* coexists with *to be going to* as a marker of future tense. *I think* as a main clause verb selecting a first-person singular subject coexists with *I think* as a parenthetical construction with great distributional freedom and with *I think* as a fixed phrase expressing speaker's epistemic attitude (Traugott 1995).

body-part expression is still in use in its original meaning, together with the more grammaticalized meaning of cardinal numeral (von Mengden 2008: 299).

Since this is an instance of grammaticalization, the regularities found cross-linguistically lead to an important observation—namely, that such consistency is unlikely to arise from some sort of featural underspecification of the original lexical item, otherwise we would expect much more heterogeneity in this process. It is instead likely that syntax itself leads the process constraining the set of possible operations that can be applied to the lexical items. This is visible in both of the hypothesized changes. Since downwards movement is not allowed, the lexical item can only move upwards; more so, being originally merged in a specifier position, its landing site is a higher specifier position (from SpecNumP to SpecDP). Serving a nominal function, the item cannot move any higher from the leftmost edge of the nominal expression, as it would exit the nominal domain. The only operation allowed at that point is the reanalysis from specifier to head position to reduce the structure to be computed. In this way, the track the whole process moves along is already traced and constrained by syntax. Assuming the universality of syntactic structures, the consistency found cross-linguistically can be accounted for.

8. Conclusion and Future Perspectives

The present work has tried to characterize the syntax underlying the process of grammaticalization of the Bulgarian numeral *edin* 'one', which already is in an advanced stage of development. Its base function is that of a cardinal numeral, quantifying the referent it is combined with. Furthermore, it also functions as a well-established specificity marker, which identifies (instead of quantifying) the NP referent it co-occurs with. *Edin* has also entered the last stage of grammaticalization, corresponding to that of the indefinite article, as it may be used non-referentially in generic contexts (cf. Geist 2013). Here I analyzed these three different functions, assigning to each of them a different structural position.

I argue that cardinal *edin* is merged in the specifier of NumP, a functional projection below the DP. Its specifier status is mainly suggested by (i) the adjectival Concord with the head N; (ii) the lack of selectional properties (e.g., the impossibility to select for the *brojna forma* shows that *edin* can only Concord with N just like APs); (iii) the existence of idiomatic expressions such as *edin i sâšt* 'one and the same' in which *edin* is coordinated with an AP; and (iv) the impossibility of being crossed by an AP.

In its function as a specificity marker, *edin* is instead hosted in SpecDP, which is the site of referential elements and the locus in which the interpretation of the whole nominal expression takes place at LF (cf. Giusti 2002). That position is supported by a strong parallelism between specific *edin* and demonstratives, which are elements argued to occupy SpecDP (cf. Brugé 2002).

Interestingly, the licensing conditions for the use of specific *edin* are the same as those which license the occurrence of “indefinite” *this* (cf. Ionin 2006), which also hold for Bulgarian “indefinite” *tozi* ‘this’.

I propose that when *edin* is used non-referentially in generic contexts, it occurs in the head D and spells out abstract Case features of the nominal expression. This parallelism is also suggested by the possibility of paraphrasing the sentences containing non-referential *edin* by substituting it with the definite article. It also follows the reanalysis from SpecDP to D that happened in the development of the definite article from demonstratives in Romance (cf. Giusti 2001).

The overall model has the advantage of describing a complete linguistic cycle which is driven by an Economy Principle imposing the minimization of interpretable features in the derivation (cf. van Gelderen 2013). The cardinal *edin* starts out in SpecNumP. Specific *edin*, however, is merged in SpecDP, applying the Late Merge Principle (van Gelderen 2011: 14), according to which the direct merging of a phrase in a higher structural position is preferred over its remerge from a lower to a higher position.⁴³ Subsequently, the Head Preference Principle (van Gelderen 2011: 13) is applied, and the specifier of the DP is reanalyzed as the head D.

Another important advantage of the model is that it allows us to account for the cross-linguistic consistency of the grammaticalization of ‘one’, assuming that syntax guides the whole process by constraining the possible operation to be applied to the lexical item undergoing this change. Assuming that the syntactic backbone is universal, the homogeneity in the stages of grammaticalization of ‘one’ stems from the same constraints imposed by syntax on the possible set of operations to be applied.

As always, there are some open issues which will need to be tackled in the future. One particular open question pertains to when the changes in the status of *edin* took place. This issue urges the investigation of corpora from Old to Modern Bulgarian to characterize this process of grammaticalization from a diachronic point of view. Moreover, another aim of future research is the possible extension and verification of this model with regard to other languages. Since the process of grammaticalization of the numeral ‘one’ is very common and consistent (the order of stages seems to be homogeneous, cf. Heine 1997) among all the natural languages, this model is likely to have cross-linguistic validity. More research is needed to confirm this claim, which I, however, deem promising. A final issue not faced here is a thorough exploration of the syntax of the plural ‘one’ *edni*. Future research will need to investigate to what

⁴³ An anonymous reviewer objects that this runs counter to the minimalist “Merge-over-Move requirement” (cf. Chomsky 1995). However, LMP does not contradict Merge-over-Move; it just favors the shorter derivation, preferring the Merge option to the combination of Merge + Move.

extent its syntax may overlap with that of the singular *edin*, and where their differences may be located in a semantic-syntactic perspective.

Sources

Bulgarian National Corpus. (2001–) Department of Computational Linguistics and the Department of Bulgarian Lexicology and Lexicography, Institute for Bulgarian Language. Available at: <http://search.dcl.bas.bg/>; also available at <http://www.sketchengine.eu>.

References

- Abney, Steven P. (1987) *The English noun phrase in its sentential aspect*. Ph.D. dissertation, MIT.
- Aijmer, Karin. (1997) “*I think* — an English modal particle”. Toril Swan and Olaf J. Westvik, eds. *Modality in Germanic languages*. Berlin/New York: De Gruyter Mouton, 1–48. DOI 10.1515/9783110889932.1.
- Alexander, Ronelle. (2000) *Intensive Bulgarian: A textbook and reference grammar*. Vols. 1 and 2. Madison: University of Wisconsin Press.
- Benacchio, Rosanna. (2018) “Slavic-Romance linguistic contact revisited: The grammaticalization of the indefinite article in the Slovene dialects of Friuli”. Thede Kahl, Iliana Krapova, and Giuseppina Turano, eds. *Balkan and South Slavic enclaves in Italy: Languages, dialects and identities*. Cambridge: Cambridge Scholar Publishing, 204–14.
- Bernstein, Judy B. (1997) “Demonstratives and reinforcers in Romance and Germanic languages”. *Lingua* 102(2–3): 87–113. DOI 10.1016/S0024-3841(96)00046-0.
- Biberauer, Theresa and Marc Richards. (2006) “True optionality: When the grammar doesn’t mind”. Cedric Boeckx, ed. *Minimalist essays*. Amsterdam: John Benjamins Publishing Company, 35–67. [Linguistik Aktuell/Linguistics Today, 91.] DOI 10.1075/la.91.08bib.
- Breu, Walter. (2012) “The grammaticalization of an indefinite article in Slavic micro-languages”. Björn Wiemer, Bernhard Wälchli, and Björn Hansen, eds. *Grammatical replication and borrowability in language contact*. Berlin/Boston: De Gruyter Mouton, 275–322. [Trends in Linguistics. Studies and Monographs, 242.] DOI 10.1515/9783110271973.
- Brugé, Laura. (2002) “The positions of demonstratives in the extended nominal projection”. Guglielmo Cinque, ed. *Functional structure in DP and IP: The cartography of syntactic structures*. Vol. 1. New York: Oxford University Press, 15–53. [Oxford Studies in Comparative Syntax.]
- Campbell, Richard. (1996) “Specificity operators in SpecDP”. *Studia Linguistica* 50(2): 161–88. DOI 10.1111/j.1467-9582.1996.tb00348.x.

- Cardinaletti, Anna and Michal Starke. (1999) "The typology of structural deficiency: A case study of the three classes of pronouns". Henk van Riemsdijk, ed. *Clitics in the languages of Europe*. Vol. 5. Berlin: Mouton de Gruyter, 145–233. [Empirical Approaches to Language Typology/Eurotyp 20–5.]
- Caruso, Đurđica Željka. (2012) *The syntax of nominal expressions in articleless languages: A split DP-analysis of Croatian nouns*. Ph.D. dissertation, University of Stuttgart.
- . (2016) "A split DP-analysis of Croatian noun phrases". *Jezikoslovlje* 17(1–2): 23–45.
- Chomsky, Noam. (1995) *The Minimalist program*. Cambridge, MA: MIT Press.
- Cinque, Guglielmo. (1994) "On the evidence for partial N-movement in the Romance DP". Guglielmo Cinque, Jan Koster, Jean-Yves Pollock, Luigi Rizzi, and Rafaella Zanuttini, eds. *Paths towards Universal Grammar: Studies in Honor of Richard S. Kayne*. Washington, D.C.: Georgetown University Press, 85–110.
- . (2005) "Deriving Greenberg's Universal 20 and its exceptions". *Linguistic inquiry* 36(3): 315–32. DOI 10.1162/0024389054396917.
- Corbett, Greville G. (1978) "Numerous squishes and squishy numerals in Slavonic". *International review of Slavic linguistics* 3(1–2): 34–73.
- Crisma, Paola. (2015) "The 'indefinite article' from cardinal to operator to expletive". Chiara Gianollo, Agnes Jäger, and Doris Penka, eds. *Language change at the syntax-semantics interface*. Berlin: Mouton de Gruyter, 125–51. [Trends in Linguistics. Studies and Monographs, 278.]
- Diewald, Gabriele. (2011) "Pragmaticalization (defined) as grammaticalization of discourse functions". *Linguistics* 49(2): 365–90. DOI 10.1515/ling.2011.011.
- Dimitrova-Vulchanova, Mila and Giuliana Giusti. (1998) "Fragments of Balkan nominal structure". *University of Venice Working Papers in Linguistics* 8: 141–72.
- Dimitrova-Vulchanova, Mila and Olga Tomić. (2009) "The structure of the Bulgarian and Macedonian nominal expression: Introduction". Mila Dimitrova-Vulchanova and Olga Tomić, eds. *Investigations in the Bulgarian and Macedonian nominal expression*. Trondheim: Tapir Academic Press, 1–23.
- Fodor, Janet Dean and Ivan A. Sag. (1982) "Referential and quantificational indefinites". *Linguistics and philosophy* 5(3): 355–98.
- Franks, Steven. (1994) "Parametric properties of numeral phrases in Slavic". *Natural language & linguistic theory* 12(4): 597–674. DOI 10.1007/BF00992929.
- . (2018) "A Bulgarian solution to the Slavic Q questions?" Steven Franks, Vrinda Chidambaram, Brian Joseph, and Iliana Krapova, eds. *Katerino Mome: Studies in Bulgarian morphosyntax in honor of Catherine Rudin*. Bloomington, IN: Slavica Publishers, 93–120.
- . (2020) *Microvariation in the South Slavic noun phrase*. Bloomington, IN: Slavica Publishers.

- Friedman, Victor A. (1976) "The question of a Bulgarian indefinite article". Thomas Butler, ed. *Bulgaria: Past and present*. Columbus: American Association for the Advancement of Slavic Studies, 334–39.
- Geist, Ljudmila. (2013) "Bulgarian *edin*: The rise of an indefinite article". Uwe Junghanns, Dorothee Fehrmann, Denisa Lenertová, and Hagen Pitsch, eds. *Formal Description of Slavic Languages: The Ninth Conference. Proceedings of FDSL 9*. Frankfurt am Main: Peter Lang, 125–48.
- Giorgi, Alessandra. (2009) "Toward a syntax of the subjunctive mood". *Lingua* 119(12): 1837–58. DOI 10.1016/j.lingua.2008.11.008.
- . (2012) "The theory of syntax and the representation of indexicality". Laura Brugé, Anna Cardinaletti, Giuliana Giusti, Nicola Munaro, and Cecilia Poletto, eds. *Functional heads: The cartography of syntactic structures*. Vol. 7. New York: Oxford University Press, 42–54. DOI 10.1093/acprof:oso/9780199746736.003.0003.
- Giusti, Giuliana. (1994) "Heads and modifiers among determiners: Evidence from Rumanian". Guglielmo Cinque and Giuliana Giusti, eds. *Advances in Roumanian linguistics*. Amsterdam: John Benjamins, 103–25. [Linguistik Aktuell/Linguistics Today, 10.]
- . (1995) "A unified structural representation of abstract and morphological case". Hubert Haider, Susan Olsen, and Sten Vikner, eds. *Studies in comparative Germanic syntax*. Dordrecht: Springer, 77–93. [Studies in Natural Language and Linguistic Theory, 31.] DOI 10.1007/978-94-015-8416-6.
- . (1997) "The categorial status of determiners". Liliane M. V. Haegeman, ed. *The new comparative syntax*. London/New York: Longman, 95–123.
- . (2001) "The rise of a functional category: From Latin ILLE to the Romance article and personal pronoun". Guglielmo Cinque and Giampaolo Salvi, eds. *Current studies in Italian syntax: Essays offered to Lorenzo Renzi*. Amsterdam: North-Holland, 157–72. [North-Holland Linguistic Series: Linguistic Variations, 59.]
- . (2002) "The functional structure of noun phrases. A bare phrase structure approach". Guglielmo Cinque, ed. *Functional structure in DP and IP: The cartography of syntactic structures*. Vol. 1. Oxford: Oxford University Press, 54–90.
- . (2006) "Parallels in clausal and nominal periphery". Mara Frascarelli, ed. *Phases of interpretation*. Berlin/New York: Mouton de Gruyter, 163–84. [Studies in Generative Grammar, 91.]
- . (2008) "Agreement and concord in nominal expressions". Cécile de Cat and Katherine Demuth, eds. *The Bantu–Romance connection: A comparative investigation of verbal agreement, DPs, and information structure*. Amsterdam: John Benjamins Publishing Company, 201–37. [Linguistik Aktuell/Linguistics Today, 131.] DOI 10.1075/la.131.12giu.
- . (2011) "On concord and projection". *Bucharest Working Papers in Linguistics* 13(1): 103–24.

- Giusti, Giuliana. (2015) *Nominal syntax at the interfaces: A comparative analysis of languages with articles*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Giusti, Giuliana and Mila Dimitrova-Vulchanova. (1996) "Quantified noun phrase structure in Bulgarian". Jindřich Toman, ed. *Formal Approaches to Slavic Linguistics: The College Park meeting, 1994*. Ann Arbor: Michigan Slavic Publications, 123–44.
- Giusti, Giuliana and Nedžad Leko. (2005) "The categorial status of quantity expressions". Nedžad Leko, ed. *Lingvistički vidici*. Sarajevo: Međunarodni Forum, 121–83. [Forum Bosnae, 34.]
- Givón, Talmy. (1981) "On the development of the numeral 'one' as an indefinite marker". *Folia Linguistica Historica* 2(1): 35–54.
- Gorishneva, Elena. (2013) "Bare vs. non-bare nouns. Two kinds of indefinites in Bulgarian". Johannes Kabatek and Albert Wall, eds. *New perspectives on bare noun phrases in Romance and beyond*. Amsterdam: John Benjamins Publishing Company, 301–28. [Studies in Language Companion Series, 141.] DOI 10.1075/slcs.141.
- Grimshaw, Jane. (1991) "Extended projections". MS., Brandeis University.
- Heine, Bernd. (1997) *Cognitive foundations of grammar*. New York: Oxford University Press.
- Heine, Bernd and Tania Kuteva. (2006) *The changing languages of Europe*. Oxford/New York: Oxford University Press.
- Horn, Laurence R. (1972) *On the semantic properties of logical operators in English*. Ph.D. dissertation, University of California-Los Angeles.
- Hwaszcz, Krzysztof and Hanna Kędzierska. (2018) "The rise of an indefinite article in Polish: An appraisal of its grammaticalisation stage (part 1)". *Studies in Polish linguistics* 13(2): 93–121. DOI 10.4467/23005920SPL.18.005.8744.
- Ionin, Tania. (2006) "This is definitely specific: Specificity and definiteness in article systems". *Natural language semantics* 14(2): 175–234. DOI 10.1007/s11050-005-5255-9.
- . (2013) "Pragmatic variation among specificity markers". Cornelia Ebert and Stefan Hinterwimmer, eds. *Different kinds of specificity across languages*. Dordrecht: Springer Netherlands, 75–103. [Studies in Linguistics and Philosophy, 92.] DOI 10.1007/978-94-007-5310-5.
- Ionin, Tania and Ora Matushansky. (2018) *Cardinals: The syntax and semantics of cardinal-containing expressions*. Cambridge, MA: MIT Press.
- Ivanova, E. Ju. and S. Koval'. (1994) "Bolgarskoe EDIN s točki zrenija referencial'nogo analiza" [Bulgarian EDIN from the point of view of referential analysis]. *Vestnik Sankt-Peterburgskogo Universiteta* 23: 58–64.
- Ivančev, Svetomir. (1957) "Nabljudenija vârxu upotrebata na člena v bâlgarskija ezik" [Observations on the use of the article in the Bulgarian language]. *Bâlgarski ezik* 7(6): 499–529.

- Jespersen, Otto. (1917) *Negation in English and other languages*. Copenhagen: A. F. Høst.
- Joseph, Brian D. (2011) "Grammaticalization: A general critique". Heiko Narrog and Bernd Heine, eds. *The Oxford handbook of grammaticalization*. Oxford: Oxford University Press, 193–205.
- Kilgarrieff, Adam, Vit Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vit Suchomel. (2014) "The Sketch Engine: Ten years on". *Lexicography* 1(1): 7–36.
- Kilgarrieff, Adam, Pavel Rychlý, Pavel Smrž, and David Tugwell. (2004) "The Sketch Engine". Technical Report ITRI-04-08, Information Technology Research Institute, University of Brighton.
- Krifka, Manfred, Francis J. Pellettier, Gregory N. Carlson, Alice ter Meulen, Gennaro Chierchia, and Godehard Link. (1995) "Genericity: An introduction". Gregory N. Carlson and Francis J. Pellettier, eds. *The generic book*. Chicago: University of Chicago Press, 1–124.
- Langacker, Ronald W. (1990) "Subjectification". *Cognitive linguistics* 1(1): 5–38.
- Leafgren, John. (2011) *A concise Bulgarian grammar*. Durham, NC: Reference Grammar Network, SEELRC, Duke University.
- Lehmann, Christian. (1995/1982) *Thoughts on grammaticalization*. Munich: LINCOM Europa. [LINCOM Studies in Theoretical Linguistics, 1.]
- Longobardi, Giuseppe. (1994) "Reference and proper names: A theory of N-movement in syntax and Logical Form". *Linguistic inquiry* 25(4): 609–65.
- Mangiulea, Mariana. (1987) "A contrastive study of един as a marker for indefinite determination in Bulgarian and the indefinite article *un* in Romanian". *Revue Roumaine de Linguistique* 32(4): 401–10.
- Marovska, Vera. (2017) "Morfologichnata kategoriya opredelenost: Neopredelenost v savremenniya balgarski ezik" [The morphological category definiteness: Indefiniteness in contemporary Bulgarian language]. *Paisii Hilendarski University of Plovdiv Research Papers* 55(1): 7–27.
- Maslov, Jurij S. (1982) *Gramatika na bălgarskija ezik* [Grammar of the Bulgarian language]. Sofia: Nauka i izkustvo.
- Nattinger, James R. (1980) "A lexical phrase grammar for ESL". *TESOL quarterly* 14(3): 337–44. DOI 10.2307/3586598.
- Nicolova, Ruselina. (2017) *Bulgarian grammar*. Berlin: Frank & Timme.
- Pancheva, Roumyana. (2018) "How many flowers! So many colors! Number marking in cardinality exclamatives in Bulgarian". Steven Franks, Vrinda Chidambaram, Brian D. Joseph, and Iliana Krapova, eds. *Katerino Mome. Studies in Bulgarian morphosyntax in honor of Catherine Rudin*. Bloomington, IN: Slavia Publishers, 197–234.
- Rappaport, Gilbert C. (2003) "Case syncretism, features, and the morphosyntax of Polish numeral phrases". *Generative linguistics in Poland* 5: 123–37.

- Rizzi, Luigi. (1997) "The fine structure of the left periphery". Liliane Haegeman, ed. *Elements of grammar: Handbook in generative syntax*. Dordrecht: Springer, 281–337. DOI 10.1007/978-94-011-5420-8_7.
- Roberts, Ian and Anna Roussou. (2003) *Syntactic change: A Minimalist approach to grammaticalization*. New York: Cambridge University Press. [Cambridge Studies in Linguistics, 100.]
- Rudin, Catherine. (2019) "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.
- Rutkowski, Paweł. (2007) "Grammaticalization in the nominal domain: The case of Polish cardinals". Blake H. Rodgers, ed. *LSO working papers in linguistics: Proceedings of WIGL 2006*. Madison: Department of Linguistics, University of Wisconsin-Madison, 89–102.
- Schroeder, Christoph. (2006) "Articles and article systems in some areas of Europe". Giuliano Bernini and Marcia L. Schwartz, eds. *Pragmatic organization of discourse in the languages of Europe*. Vol. 8. Berlin/New York: Mouton de Gruyter, 545–611. [Empirical Approaches to Language Typology/Eurotyp 20-8.]
- Stoevski, Andrei. (2019) "Status and uses of *edin* 'one' in contemporary Bulgarian". Dimităr Veselinov, ed. *Godišnik na Sofijskija universitet "Sv. Kliment Oхridski": Fakultet po klasičeski i novi filologii* [Annual of Sofia University "St. Kliment Ohridski": Faculty of classical and modern philology]. Vol. 112. Sofia: St. Kliment Ohridski University Press, 177–219.
- Traugott, Elizabeth C. (1989) "On the rise of epistemic meanings in English: An example of subjectification in semantic change". *Language* 65(1): 31–55. DOI 10.2307/414841.
- . (1995) "Subjectification in grammaticalization". Dieter Stein and Susan Wright, eds. *Subjectivity and subjectivisation: Linguistic perspectives*. New York: Cambridge University Press, 31–54.
- Trousdale, Graeme and Muriel Norde. (2013) "Degrammaticalization and constructionalization: Two case studies". *Language sciences* 36: 32–46. DOI 10.1016/j.langsci.2012.03.018.
- van Gelderen, Elly. (2008) "Linguistic cycles and Economy Principle: The role of Universal Grammar in language change". Thórhallur Eythórsson, ed. *Grammatical change and linguistic theory: The Rosendal papers*. Amsterdam: John Benjamins, 245–64. [Linguistik Aktuell/Linguistics Today, 113.]
- . (2011) *The linguistic cycle: Language change and the language faculty*. New York: Oxford University Press.
- . (2013) "The linguistic cycle and the language faculty". *Language and linguistics compass* 7(4): 233–50. DOI 10.1111/lnc3.12017.

- von Mengden, Ferdinand. (2008) "The grammaticalization cline of cardinal numerals and numeral systems". María José López-Couso and Elena Seoane, eds. *Rethinking Grammaticalization: New perspectives*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 289–308. [Typological Studies in Language, 76.]
- Weiss, Daniel. (2004) "The rise of an indefinite article: The case of Macedonian *eden*". Walter Bisang, Nikolaus Himmelmann, and Björn Wiemer, eds. *What makes Grammaticalization? A look from its fringes and its components*. Berlin/New York: Mouton de Gruyter, 139–65. [Trends in Linguistics. Studies and Monographs, 158.]

Luca Molinari

University of Warsaw

Warsaw, Poland

l.molinari@uw.edu.pl

Ca' Foscari University of Venice

Venice, Italy

luca.molinari@unive.it

Threatening in Russian with or without *sja*: *Grozit'* vs. *grozit'sja**

Tore Nessel and Anastasia Makarova

Abstract: This article explores the two verbs, *grozit'* and *grozit'sja*, which can both be translated as 'threaten'. We adopt a "local" approach and offer a thorough analysis of corpus data, which indicates that the two verbs, although they share a number of properties, are semantically and syntactically distinct. We show that the two verbs collocate with different parts of speech and tend to occur in different syntactic constructions. *Grozit'sja* is typically used with regard to interactions between two persons, while *grozit'* has a wider range of uses. This tendency has become more pronounced over time. As for the meaning of the verbs, *grozit'sja* tends to express verbal threats, while *grozit'* often conveys non-verbal threats. On a more theoretical level, our study contributes to our understanding of the morpheme *sja*. While labels like "reflexive", "middle", and "passive" are helpful as far as they go, we demonstrate how detailed studies of individual verb pairs (a "local" approach) may shed light on the complex syntactic and semantic properties of *sja*. On the methodological level, our study underscores the value of corpus data for the study of *sja*, both data from large internet corpora such as the Araneum Russicum Russicum Maius and the Russian National Corpus (RNC). While the former corpus enables us to identify general tendencies through collocations and semantic vectors, a smaller curated corpus like the RNC is suitable for detailed analysis of semantic and syntactic properties.

1. Introduction: The Problem

Notorious for its polyfunctionality, the morpheme *sja* represents a classic descriptive and theoretical problem in Russian linguistics. What is the meaning of *sja*? What is the semantic and syntactic effect of adding *sja* to a verb? As is well known, *sja* is attested as a marker of middle (or reflexive) voice

* We would like to express our gratitude to our colleagues in the CLEAR (Cognitive Linguistics: Empirical Approaches to Russian) research group at UiT The Arctic University of Norway. Thanks to Sergey Say for discussing an earlier version of the paper with us, and to two anonymous reviewers and the editors of *JSL* for detailed and helpful comments.

in (1) and passive voice in (2), as well as in a number of related functions, for which a wide variety of classifications and terminologies exist.¹

- | | | | | | |
|-----|-------------|--------|-------------|-------------------|-------|
| (1) | Francuzskij | korol' | Ljudovik XI | my-l-Ø-sja | pjat' |
| | French | king | Louis XI | wash-PST-SG-REFL | five |
| | raz-Ø | v | god-Ø. | | |
| | time-GEN.PL | in | year-ACC.SG | | |
- 'The French king Louis XI washed five times a year.'

- | | | | | | |
|-----|---------------|--------------------|------------------|---------------|----|
| (2) | Kak | ob"jasni-t' | istori-ju | pojaveni-ja | na |
| | how | explain-INF | history-ACC | emergence-GEN | on |
| | territori-i | Rossi-i | nemetsk-ix | kirx-Ø, | |
| | territory-LOC | Russia-GEN | German-GEN.PL | church-GEN.PL | |
| | kotor-ye | stoi-l-i-s' | tevton-ami [...] | | |
| | which-NOM.PL | build-PST-PL-REFL | teutons-INS.PL | | |
- 'How can we explain the emergence of German churches on Russian territory, that were built by the Teutons [...].'

A note on terminology is necessary. Many researchers refer to examples like (1) as "reflexive" and verbs like *grozit'sja* as "reflexive verbs", but following Kemmer (1993) and Enger and Nessel (1998), we prefer the term "middle voice" for examples with *sja*, thus reserving the term "reflexive" for sentences with the pronoun *sebjja* (e.g., *nenavidet' sebjja* 'hate oneself'). Although the term "middle voice" has not been used so much in Russian and Slavic linguistics, we find it helpful since Russian has a grammatical distinction between *sebjja* and *sja*, for which we can use the terms "reflexive" and "middle", respectively. This usage also comes with the advantage that it is in harmony with typological works on voice distinctions (e.g., Kemmer 1993).

Traditionally, *sja* has been described in terms of a list or, especially in cognitive and functionally-oriented linguistics, a network of related meanings or functions (see, for example, Geniušienė 1987; Enger and Nessel 1998; Goto and Say 2009; Kyröläinen 2013). In order to shed light on this list or network, two approaches are conceivable. A "global" approach involves investigating and classifying a wide variety of verbs, while what we may call a "local" approach offers in-depth analyses of individual verbs. The two approaches are complementary, and in actual practice, most researchers who offer "global" analyses of the system of *sja* as a whole also, to some extent, provide "local"

¹ All numbered examples are from the Russian National Corpus (www.ruscorpora.ru). For the convenience of the reader, the relevant verb or construction is boldfaced. Our database is available in TROLLing (The Tromsø Repository of Language and Linguistics); see Makarova and Nessel 2022.

descriptions of individual verbs (e.g., Israeli 1997; Knjazev 2007; Goto and Say 2009).

In the present study, we adopt a “local” approach where we focus on the two near synonyms *grozit'* and *grozit'sja*, both of which can be translated as ‘threaten’ (Glovinskaja 2004a, 2004b).² The two verbs can be attested in very similar syntactic environments. In (3) and (4), for instance, both verbs combine with a nominative subject representing the “threatener” (the person who carries out the threat), a noun phrase in the dative representing the “threatenees” (the person who is threatened), and an infinitive complement representing the action the subject threatens to carry out:³

- (3) Skol'ko raz-Ø Carevskij i Vevers
 how.many time-GEN.PL Carevskij and Vevers
 grozi-l-i mne sostavi-t' protokol o mo-ix
 threaten-PST-PL I.DAT compile-INF protocol about my-LOC.PL
 popyt-k-ax “diskreditirova-t' rukovodstv-o
 attempt-LOC.PL discredit-INF leadership-ACC
 obkom-a [...].
 regional.committee-GEN.SG

‘How many times didn’t Carevskij and Vevers threaten me to report my attempts to “discredit the leadership of the regional committee”.’

- (4) [V]y časten'ko **grozi-l-i-s'** Čebakov-u **ujti**
 you.PL often threaten-PST-PL-REFL Čebakov-DAT leave-INF
 k svo-emu professor-u [...].
 to own-DAT.SG.M professor-DAT.SG

‘[Y]ou often threatened Čebakov to go to your professor [...].’

² It is worth pointing out that Russian has a number of verb pairs with and without *sja*, which deserve closer analysis. Examples include *dymit'* – *dymit'sja* ‘smoke’, *kružit'* – *kružit'sja* ‘spin’, *rešit'* – *rešit'sja* ‘decide’, and *xvastat'* – *xvastat'sja* ‘boast’ (cf. Israeli 1997: 95–107; Gerritsen 1990: 95–97). For each of these pairs, it is difficult to pinpoint the exact semantic contribution of *sja*.

³ Notice that “threatener” and “threatenees” do not have to be persons. For instance, in *Mne grozila smertnaja kazn'* (lit.) ‘Death penalty threatened me’, we analyze *smertnaja kazn'* ‘death penalty’ as the “threatener” (see also Section 5 below). We only analyze constituents that are overtly expressed in the examples. We would like to emphasize that “threatener” and “threatenees” are invariable semantic categories that can be realized as different syntactic functions (subject, object, etc.). While it would be interesting to carry out a systematic analysis of the relationship between “threatener” and “threatenees” on the one hand and syntactic functions on the other, such an analysis is beyond the scope of the present analysis.

In examples like (3) and (4), *grozit'* and *grozit'sja* may be used interchangeably without clear semantic differences. We must therefore ask: What is the meaning of *sja* in *grozit'sja*? What is the effect of adding *sja* to *grozit'*? While several researchers have provided insightful analyses (e.g., Gerritsen 1990; Israeli 1997), we are not aware of extensive investigations of data from large electronic corpora, using the methodologies of contemporary corpus linguistics. The present study aims at filling this knowledge gap.

Besides offering an analysis of *grozit'* and *grozit'sja* that has implications for our understanding of *sja* in general, we address the culturally and linguistically important concept of “threat” that has received considerable attention in general linguistics in recent years. Cognitive and functionally-oriented linguists have discussed the verbs for threatening in English, Dutch, and Spanish (cf., for example, Langacker 1999; Verhagen 1995; Cornillie 2004). Examples like *The incident threatened to ruin his chances* (Verhagen 1995: 111) are argued to involve a high degree of “subjectification”, whereby the likelihood of the relevant event (e.g., *to ruin his chances*) receives a positive or negative evaluation by the speaker (Cornillie 2004).

Experts on grammaticalization have been interested in verbs for threatening since they represent a grammaticalization path from examples like (3) and (4), where a person promises to harm another person, to more abstract examples such as *The Australian dollar threatens to fall below 72 cents* (Narrog and Heine 2021: 32; see also Heine and Miyashita 2007, 2008). Although subjectification and grammaticalization are not central topics of the present study, we note that Russian is of particular interest for linguistic investigations of threats since Russian has more than one morphologically related verb for ‘threaten’. In addition to *grozit'* and *grozit'sja*, Russian also has the prefixed imperfective verb *ugrožat'*, as well as a number of prefixed perfective verbs, such as *prigrozit'*, which all can be translated as *threaten*. In the present study, we limit ourselves to *grozit'* and *grozit'sja*, which are relevant for the study of *sja*.

The contribution of our study can be summarized as follows. First, we show that *grozit'* and *grozit'sja*, although they show some degree of overlap, are syntactically and semantically distinct. Second, our study illustrates the value of a “local approach” to *sja*. While simple labels like “middle voice” and “passive” are useful as far as they go, we also need detailed analyses of individual verbs in order to pinpoint all the idiosyncratic and unpredictable properties of *sja*. Third, on the methodological level, our analysis indicates the usefulness of investigating semantic vectors and collocations in large internet corpora. However, at the same time, we show that detailed analysis of individual examples from curated and balanced corpora is also required. Last but not least, our analysis demonstrates that Russian, like other European languages, has abstract examples that deviate from the prototypical situation where one person promises to do harm to another person. Interestingly,

this mainly applies to *grozit'*, while *grozit'sja* is more likely to be used about prototypical threats.

Our argument is structured as follows. Sections 2 and 3 are devoted to semantic vectors and collocations in a large internet corpus (Araneum Russicum Russicum Maius).⁴ In sections 4 and 5, we turn to data from the Russian National Corpus and consider argument structure constructions. Sections 6 and 7 concern the meaning of the two verbs under scrutiny, before we turn to the meaning and functions of *sja* in Section 8. Section 9 summarizes our findings.

2. Semantic Vectors: How Similar Are *grozit'* and *grozit'sja*?

As a first step in our attempt at teasing apart the meanings and functions of *grozit'* and *grozit'sja*, we use semantic vectors (word embeddings), a method that has been gaining importance in corpus studies in recent years. As we will see, *grozit'* and *grozit'sja* do not come out as close relatives, but both are indirectly related through their common relative *ugrožat'*, which also means 'threaten'.

The idea behind semantic vectors is the Distributional Hypothesis that words with similar meanings tend to occur in similar contexts. If you happen not to know the word *sriracha* but notice that it shows up in texts together with *hamburger* in much the same way as, say, *ketchup*, *aioli*, *mayonnaise*, and *béarnaise*, you might correctly guess that *sriracha* is a sauce that goes well with hamburgers. While the Distributional Hypothesis goes back at least to the 1950s (Joos 1950; Harris 1954; Firth 1957), it was only with the advent of large electronic corpora that it was possible to make real use of it. Combined with large corpora, semantic vectors offer enormous power to Natural Language Processing, as pointed out by Jurafsky and Martin (2024). It is possible to calculate a vector for each word based on all the contexts where it is attested in a corpus. The vector of each word can be represented as a point in a multidimensional space, where similar words are located close to each other.

The Araneum Russicum Russicum Maius corpus, a large internet corpus containing 1.2 billion Russian word tokens, includes a function that enables us to investigate the similarity of words by means of semantic vectors.⁵ For each word one searches for, the corpus returns a list of the 25 most closely

⁴ The Araneum Russicum Russicum Maius corpus is available at <http://unesco.uniba.sk/>.

⁵ The function for assessing similarities among words can be found here: <https://www.juls.savba.sk/sem%C3%A4/?lang=ru&kio=lemma&visualsel=gnuplot&topn=24&wpos=&wneg=>. Our searches were carried out on 19 November 2021. A detailed discussion of the technical procedures behind the calculations of semantic vectors in the Araneum corpus is beyond the scope of the present study.

related words, where “related” means that they occur in similar contexts in the corpus. Table 1 provides the lists for *grozit'*, *grozit'sja*, as well as the third imperfective verb for ‘threaten’, *ugrožat'*. As shown, the list for *grozit'* does not contain *grozit'sja*, and the list for *grozit'sja* does not contain *grozit'*. However, both lists include *ugrožat'*, and the list of *ugrožat'* contains both *grozit'* and *grozit'sja*. In other words, the semantic vectors from the Araneum corpus indicate that *grozit'* and *grozit'sja* are related, but only indirectly through *ugrožat'*. Both *grozit'* and *grozit'sja* are related to *ugrožat'*.

Table 1. The 25 most closely related words to *grozit'*, *grozit'sja*, and *ugrožat'* based on semantic vectors from the Araneum Russicum Russicum Maius corpus

grozit'	grozit'sja	ugrožat'
0.000, <i>grozit'</i>	0.000, <i>grozit'sja</i>	0.000, <i>ugrožat'</i>
0.488, ugrožat'	0.295, <i>grozilas'</i>	0.470, <i>ugrožajuščego</i>
0.536, <i>črevatyj</i>	0.377, <i>prigrozili</i>	0.488, grozit'
0.562, <i>grozjaščee</i>	0.384, <i>prigrozit'</i>	0.519, <i>ugroza</i>
0.601, <i>grozjaščij</i>	0.406, <i>prigrozila</i>	0.520, <i>ugrožajuščij</i>
0.632, <i>grozjaščego</i>	0.510, <i>grozjas'</i>	0.559, <i>ugrožavšego</i>
0.635, <i>grozjaščix</i>	0.577, <i>poobeščat'</i>	0.609, grozjaščix
0.640, <i>grozjaščij</i>	0.611, <i>obmateril</i>	0.620, <i>prigrozit'</i>
0.644, <i>povleč'</i>	0.612, <i>vygnat'</i>	0.620, <i>šantažirovat'</i>
0.662, <i>črevatyj</i>	0.613, <i>otmestka</i>	0.626, <i>šantažirovali</i>
0.666, <i>grozivšij</i>	0.613, <i>voznamerit'sja</i>	0.645, grozit'sja
0.668, <i>obernut'sja</i>	0.615, <i>pripugnul</i>	0.646, <i>ugrožaem</i>
0.683, <i>grozivšaja</i>	0.628, <i>zasudjat</i>	0.647, grozjaščego
0.687, <i>karat'sja</i>	0.630, <i>zasudit</i>	0.655, <i>zapugival</i>
0.691, <i>neminuemyj</i>	0.636, <i>posmet'</i>	0.663, <i>zapugivat'</i>
0.693, <i>štraf</i>	0.645, ugrožat'	0.671, <i>ugrožaj</i>
0.693, <i>prigrozit'</i>	0.649, <i>šantažirovali</i>	0.675, grozjaščee
0.696, <i>sprovotsiruet</i>	0.650, <i>požaluetsja</i>	0.677, <i>obespokoit'</i>
0.702, <i>vleč'</i>	0.651, <i>naoral</i>	0.684, grozjaščij
0.712, <i>grozjaščem</i>	0.662, <i>vyšvyrnut'</i>	0.684, grozjaščij

—continued on next page—

—continued—

grozit'	grozit'sja	ugrožat'
0.714, naneslo	0.665, našalovalas'	0.685, opasat'sja
0.716, obespokoit'	0.666, zapugival	0.689, ugrožajuščem
0.716, prigrozili	0.666, šantažirovat'	0.693, ugrožajuščego
0.717, nakazyvat'sja	0.666, podgovorili	0.694, ugrožavšuju
0.717, grozivšej	0.667, nakažut	0.694, prigrozili

To summarize, our analysis of the semantic vectors from the Araneum Russicum Russicum Maius corpus strongly suggests that *grozit'* and *grozit'sja* are not complete synonyms. On the basis of corpus data, it should therefore be possible to pinpoint the differences between the two verbs—a task we turn to in the following sections.

3. Collocations

A useful function of the Araneum Russicum Russicum Maius corpus is to search for collocations, i.e., words that are likely to co-occur with *grozit'* and *grozit'sja*.⁶ As we will see, the two verbs turn out to have different profiles when it comes to collocations. This lends further support to the observation that the two verbs are not perfect synonyms.

We searched for collocations of *grozit'* and *grozit'sja* with the specification that the distance between the verb and the other word be from +1 to −1 word. In this way, we identify the words immediately preceding and following the verbs under scrutiny. The corpus offers several ways of ranking the collocations. We chose the logDice option, which is useful for data from large corpora since it does not take into account corpus size. The 50 most highly ranked collocations for *grozit'* and *grozit'sja* are listed in Table 2 on the following page.

⁶ We also checked the collocation function in CoCoCo (Collocations, Colligations, Corpora, <https://cococo.cosyco.ru>), but this tool did not return relevant results for *grozit'sja*, which is less frequent than *grozit'*.

Table 2. The 50 most highly ranked collocations for *grozit'* and *grozit'sja* from the Araneum Russicum Russicum Maius corpus⁷

<i>grozit'</i>	logDice	<i>grozit'sja</i>	logDice
štraf	8.19887	vzvintit'	6.56121
opasnost'	7.69859	vygnat'	6.31525
lišenie	7.27693	vyselit'	5.54810
nakazanie	6.91199	otomstit'	5.51982
obernut'sja	6.88363	podžeč'	5.20396
narušitel'	6.86265	otratit'	5.12540
tjurennyj	6.84390	obrušit'sja	4.61792
gibel'	6.45812	uvolit'	4.59462
pererasti	6.29104	vot-vot	4.43086
smertnyj	6.16966	sžeč'	4.35690
neminuemyj	6.04915	razorvat'	4.26847
ser'ěznyj	5.85520	pobit'	4.19426
uvol'nenie	5.78396	pererasti	4.18515
ugolovnyj	5.66432	vypisat'	4.13041
požiznennyj	5.63443	otnjat'	4.08039
pal'čik	5.60695	vykinut'	4.01772
vymiranie	5.58353	ubit'	3.97398
poterja	5.57361	zabrat'	3.84047
tjur'ma	5.52686	nakazat'	3.47399
smertel'nyj	5.51174	otobrat'	3.41720
kulak	5.45412	podat'	3.32377
beda	5.41713	zapravka	2.88311
bankrotstvo	5.28178	otmenit'	2.85241
arest	5.25167	lišit'	2.73370
katastrofa	5.25131	brošit'	2.63994
sryv	5.20595	razrušit'	2.46576
obval	5.13969	ujti	2.34544
zatjanut'sja	5.13802	sdat'	2.24015
sanktsijami	5.13228	posadit'	2.15601

—continued on next page—

⁷ Collocations are ranked according to logDice. High numbers indicate a high likelihood for a word to occur next to *grozit'*/*grozit'sja*.

<i>—continued—</i> <i>grozit'</i>	logDice	<i>grozit'sja</i>	logDice
neprijatnost'	5.11375	vyvesti	1.96016
letal'nyj	5.01836	uničtožit'	1.93767
defolt	5.00699	zakryt'	1.83435
diskvalifikatsija	5.00468	prevratit'sja	1.83407
administrativnyj	4.98508	priexat'	1.53527
čelovečstvo	4.96115	otpravit'	1.50875
isčeznovenie	4.95020	pozvonit'	1.27813
obrušeniem	4.93257	činovnik	1.27485
čem	4.91842	jandeks	1.24431
promedlenie	4.88894	zapustit'	1.23064
osložnenie	4.85271	opublikovat'	1.10860
smert'	4.84071	muž	1.04212
vot-vot	4.81574	davno	0.98723
razorenje	4.81207	tsar'	0.97841
prevratit'sja	4.80409	paren'	0.97348
razrušenje	4.77729	vvesti	0.91549
deportatsija	4.76885	peredat'	0.88388
voditel'	4.76812	ustroit'	0.85994
obrušit'sja	4.71030	otdat'	0.81448
učast'	4.65922	povysit'	0.78598
besplodie	4.59036	ostavit'	0.65668

We would like to draw attention to two facets of the lists in Table 2. First, we see that the logDice values are generally higher for *grozit'* than for *grozit'sja*. This suggests that *grozit'sja* is more flexible with regard to the contexts it occurs in, while *grozit'* may have closer ties to its collocates. Second, the two lists are quite different, which shows that the two verbs typically combine with different words. The second point becomes even clearer if we classify the collocates with regard to their parts of speech.

As shown in Figure 1 and Table 3, *grozit'* tends to combine with nouns and, to a lesser degree, adjectives, while *grozit'sja* typically co-occurs with verbs. The small category “other” in the table includes adverbs and pronouns. The differences are statistically significant with a large effect size.⁸

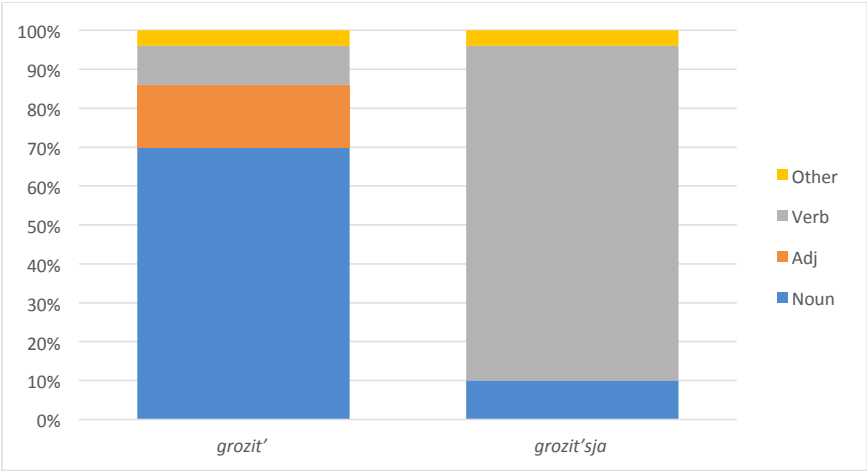


Figure 1. The 50 most highly ranked collocations for *grozit'* and *grozit'sja* sorted according to parts of speech

Table 3. The 50 most highly ranked collocations for *grozit'* and *grozit'sja* sorted according to parts of speech

	<i>grozit'</i>	<i>grozit'sja</i>
Noun	35	5
Adjective	8	0
Verb	5	43
Other	2	2

Our analysis of collocations in the Araneum Russicum Russicum Maius corpus brings us one step closer to pinpointing the differences between *grozit'* and *grozit'sja*. Knowing that the former prefers combinations with nouns,

⁸ We compared the numbers for nouns and verbs for *grozit'* and *grozit'sja*. Pearson's Chi-squared test with Yates' continuity correction (X-squared = 49.225, df = 1) returned a *p*-value = 2.282e-12. Cramer's V-value was calculated to 0.77, which indicates a large effect size.

while the latter typically collocates with verbs, we can proceed to a more detailed analysis of the constructions, in which *grozit'* and *grozit'sja* occur.

4. Constructions: Argument Structure

In order to get a clearer picture of the constructions of *grozit'* and *grozit'sja*, we created a database with examples from the Russian National Corpus.⁹ This corpus is smaller than the Araneum Russicum Russicum Maius corpus, but it is curated and provides considerable metadata for each example, thus facilitating in-depth analysis. Our data confirm the observations from the two previous sections that *grozit'* and *grozit'sja* show different behavior (see also Glovinskaja 2004a, 2004b for discussion).

Our database was constructed as follows. We searched for both verbs in five time periods: 1800–1849, 1850–1899, 1900–1949, 1950–1999, 2000–present. For each period, we made a random sample of 50 examples for each verb. In order to avoid biased samples, we only included one example for each author. *Grozit'sja* is less frequent than *grozit'*, and for the 1800–1849 period, we were only able to include 23 examples in the database. All in all, the database thus contains 473 examples—250 for *grozit'* and 223 for *grozit'sja*. The examples were manually annotated for their syntactic constructions, as well as several other parameters, which we will come back to in later sections.

For the purposes of our analysis, we distinguish between five constructions. A frequent pattern is for the verbs to combine with a nominative subject, an argument in the dative, and additional constituents. We refer to this construction as “NomVDat+”:¹⁰

- (5) I oni zna-l-i: u babuški Dženni
 and they know-PST-PL at grandmother Jenny
 im **ne** **groz-jat** poučeni-ja.
 they.DAT not threaten-3PL homily-NOM.PL
 ‘And they knew: at grandmother Jenny’s place they would not be
 threatened with any homilies.’

⁹ We used the main subcorpus of the Russian National Corpus, which contains approximately 330 million words. Corpus searches were carried out on 23 September 2021.

¹⁰ In the abbreviations for the constructions, “V” stands for the verb *grozit'* or *grozit'sja*. The + sign indicates the possibility of additional arguments in the construction. Notice that the order of constituents has not been taken into consideration. In (5) and (6), for instance, the dative argument occurs in different positions, but we analyze both examples as the same construction.

- (6) Posledn-juju tirad-u on proiznės-Ø s
 last-ACC.SG.F rant-ACC.SG he pronounce-PST.SG.M with
 bol's-oj sil-oj, budto **groz-ja-s'** **komu-to.**
 big-INS.F force-INS.SG as.if threaten-CVB-REFL someone.DAT
 'The last rant he pronounced very vigorously, as if he was threatening
 someone.'

We have quite a few examples where the verb co-occurs with a nominative subject and a complement in the instrumental, as well as additional constituents. We call this construction "NomVInstr+":¹¹

- (7) Poètomu problem-y so zdorov'-em by-l-i, a
 so problem-PL with health-INS.SG be-PST-PL and
 zaboľevanie, nača-vš-ee-sja 31 ijul-ja, v
 illness start-PTCP-N.SG-REFL 31 july-GEN.SG in
 dal'nejš-em **groz-it** **ser'jėzn-ymi**
 future-LOC threaten-3SG serious-INS.PL
osložnenij-ami.
 complication-INS.PL
 'So there were some health issues, and the illness that started 31 July,
 can have serious complications.'

- (8) I ona daže **groz-it-sja** **pal'c-em.**
 and she even threaten-3SG-REFL finger-INS.SG
 'And she is even making threatening gestures with her finger.'

A combination of dative and instrumental complements is found in examples of the following type:

- (9) Ja **nič-em** i **nik-omu** **ne**
 I nothing-INS.SG and no.one-DAT.SG not
grož-u.
 threaten-PRS.1SG
 'I am not threatening anyone with anything.'

¹¹ Notice that the noun phrases in the instrumental can represent the potential consequence of the threatening situation, as in (7), or the body part or weapon that is used in the relevant situation to threaten someone, as in (8). Both types are attested for both verbs in our database.

- (10) — U, zaraz-y, — **grozi-l-a-s'** **derev'-jam**
 oh bastard-NOM.PL threaten-PST-F-REFL tree-DAT.PL
 babuška **suxon'k-im** **kulak-om.**
 grandmother dry-M.INS.SG fist-INS.SG
 'Oh, you bastards, the old woman threatened the trees with her dry
 fist.'

The previous examples involve nominal complements. However, another important construction, for which we use the label "NomVInf", involves a nominative subject and an infinitive complement:

- (11) Molčanov sta-l-Ø za nim,
 Molčanov stand-PST-M.SG behind he.INS
 vynu-l-Ø špag-u i **grozi-l-Ø** **izrubi-t'**
 take out-PST-M.SG sword-ACC and threaten-PST-M.SG cut-INF
 ego, eželi on strus-it.
 he.ACC if he act like a coward-FUT.3SG
 'Molčanov stood behind him, he pulled out his sword and threatened
 that he would cut him in pieces if he should act like a coward.'
- (12) Priš-l-i medsěstr-y i skaza-l-i, čto
 come-PST-PL nurse-NOM.PL and say-PST-PL that
 on **groz-it-sja** ix vs-ex **poubiva-t'.**
 he threaten-PRS.3SG-REFL they.ACC all-ACC.PL kill-INF
 'The nurses came and said that he was threatening to kill them all.'

It is not uncommon for *grozit'* and *grozit'sja* to occur with a subject in the nominative but no complement. Examples of this NomV construction are demonstrated in the following examples:

- (13) Ja ne **grož-u,** ja ne
 I not threaten-PRS.1SG I not
 vymoga-ju prošč'enij-a.
 extort-PRS.1SG forgiveness-GEN.SG
 'I am not threatening, neither am I extorting forgiveness.'

- (14) On ne obiza-l-Ø-sja, ne plaka-l-Ø, ne
 he not offend-PST-M-REFL not cry-PST-M not

grozi-l-Ø-sja [...].

threaten-PST-M-REFL

'He was not getting offended, did not cry or threaten [...].

Finally, we have a number of attestations where the verb co-occurs with a clausal complement in addition to the nominative subject. We refer to this construction as "NomVClause":¹²

- (15) Eščě do ot'ezd-a on v
 already before departure-GEN he in
 razgovor-ax s drug-imi **grozi-l-Ø,** **čto**
 conversation-LOC.PL with other-INS.PL threaten-PST-M that
 Griboedov-u èt-a šutka ne projd-ët
 Griboedov-DAT this-F.SG joke not pass-FUT.3SG
 darom.
 for.free

'Even before he left, in conversations with the others he was threatening that this joke would not go without consequences for Griboedov.'

- (16) Potom ja plaka-l-a i ona menja
 then I cry-PST-F and she I.ACC
 uteša-l-a, **grozi-l-a-s',** **čto** sladk-ogo ne
 comfort-PST-F threaten-PST-F-REFL that sweet-GEN.SG not
 da-st.
 give-FUT.3SG

'Then I was crying, and she was comforting me, threatening that she would not give me any sweets.'

The distribution of these constructions in our database is summarized in Figure 2 and Table 4 where the category "other constructions" includes miscellaneous types, e.g., with prepositions or participles in oblique cases. As shown, the two verbs are attested in the same constructions but nevertheless have different profiles. For *grozit'*, by far the most frequent construction is NomVDat+, which is rare for *grozit'sja*. Other frequent constructions for *grozit'*

¹² Notice that we also include examples where *grozit'* or *grozit'sja* is followed by direct speech in the NomVClause category.

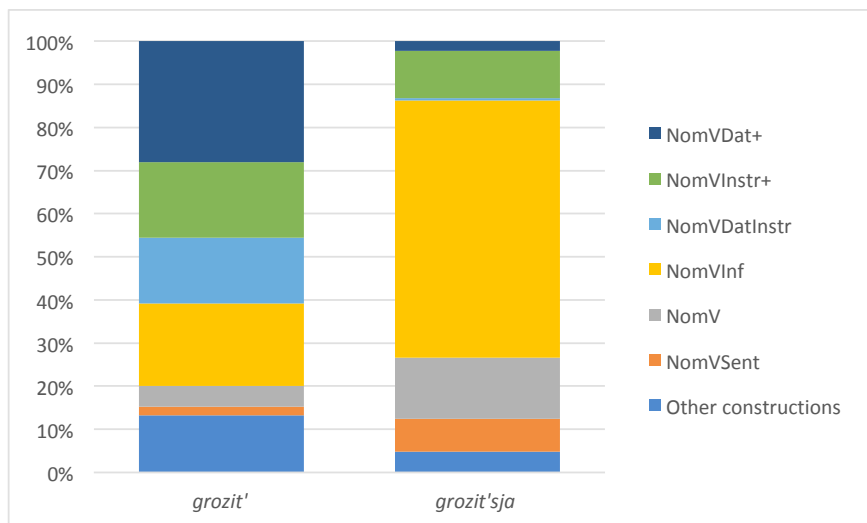


Figure 2. The distribution of constructions with *grozit'* and *grozit'sja* (data from the Russian National Corpus)

Table 4. The distribution of constructions with *grozit'* and *grozit'sja* (data from the Russian National Corpus)

	<i>grozit'</i>	<i>grozit'sja</i>
NomVDat+	70	5
NomVInstr+	44	25
NomVDatInstr	38	1
NomVInf	48	135
NomV	12	32
NomVSent	5	17
Other constructions	33	11
Total	250	223

involve complements in the instrumental or the combination of instrumental and dative complements. By contrast, the most frequent option for *grozit'sja* is the NomVInf construction, which is much less frequently attested for *grozit'*. These results square with the findings from the Araneum Russicum Russicum Maius corpus presented in the previous section, insofar as *grozit'* typically combines with nominal arguments, whereas *grozit'sja* prefers an infinitive

complement. The observed differences are statistically highly significant and show a large effect size.¹³

For the other constructions, the numbers are smaller and the differences less clear, but it is interesting to notice that the NomVInstr+ construction is more frequent for *grozit'* than for *grozit'sja*. Once again, we see that *grozit'* has the stronger affinity for nominal complements (here, an NP in the instrumental case).

Before we conclude, two methodological points deserve mention. First, we have focused on the constructions that are attested in the data. Here, we follow the usage-based approach of cognitive linguistics, where generalizations are assumed to be based on the patterns in actual language usage (Langacker 1991: 261–88 and 1999: 91–146). This methodology allows us to establish typical patterns, but we are not in a position to identify all possible constructions. Furthermore, we cannot identify which constructions are impossible.¹⁴

A second methodological point concerns the level of analysis. For the purposes of our study, we have characterized the arguments in terms of case and distinguished between nominal and clausal complements. It would be possible to create a more detailed analysis by adding, for instance, semantic roles. While this would have added another dimension to the analysis, it would yield a large number of small categories, on the basis of which no statistically robust generalizations could be made. We have therefore not added further semantic layers to our classification of constructions.

To summarize, our analysis of data from the Russian National Corpus indicates that *grozit'* and *grozit'sja* are syntactically different, insofar as they tend to occur in different constructions. While *grozit'* typically takes a nominal complement in the dative or instrumental cases, *grozit'sja* is most frequently attested with an infinitive, a fact we will return to in Section 8. We hasten to add that the observed differences are not categorical. Both verbs are attested in all the constructions we have explored in this section—but with very different frequencies.

¹³ We compared the numbers for NomVDat+, NomVInstr+, and NomVDatInstr on the one hand with the numbers for NomVInf on the other. Pearson's Chi-squared test with Yates's continuity correction (X-squared = 116.95, df = 1) returned a *p*-value < 2.2e-16. Cramer's V-value is 0.6, which represents a large effect size.

¹⁴ An anonymous reviewer points out that a dative argument is hardly compatible with a complement clause. According to him/her, examples like *On grozil emu vygnat'* 'He threatened to chase him away' without an explicit object in the embedded clause are completely unacceptable. We share the intuitions of the reviewer and agree that an investigation of such restrictions might be fruitful. However, in order to test the hypothesis of the reviewer properly, we would need an extensive survey with a large number of native speakers. Such an investigation is beyond the scope of the present study.

5. Arguments: Persons vs. Non-Persons

A prototypical threat may be characterized as a situation where one person promises to do harm to another person, as in examples (3) and (4), cited in Section 1. In other words, we are dealing with a relationship between two persons. In what follows, we show that this prototypical scenario is characteristic of *grozit'sja*, whereas *grozit'* has developed abstract meanings, following a grammaticalization path that is well known from other European languages (Heine and Miyashita 2007, 2008; Narrog and Heine 2021).

Peškovskij (1956: 119) and Gerritsen (1990: 96) have mentioned that *grozit'sja* combines with subjects that refer to persons, while *grozit'* does not have such a restriction. In order to test this hypothesis against corpus data, we distinguish between two broad categories, “persons” and “non-persons”, where the latter category includes both entities (concrete objects and abstract concepts) and events.¹⁵ Here are relevant examples with non-persons:

- (17) Za tjažk-ie prestupleni-ja ej
for serious-ACC.PL crime-ACC.PL she.DAT
grozi-l-a smertnaja kazn'.
threaten-PST-F death.ADJ penalty

‘For her serious crimes a death penalty was threatening her.’

- (18) Aprel' mesjac stoja-l-Ø v polovin-e, **dorog-i**
April month stand-PST-M in middle-LOC road-NOM.PL
grozi-l-i-s' sdela-t'-sja neproxodim-yμι.
threaten-PST-PL-REFL become-INF-REFL impassible-INS.PL

‘It was the middle of April, and the roads threatened to become impassable.’

Notice that it is not only the “threatener” that can be a “non-person”. In the following examples, the “threatenee” is not a person. In (19), Africa is threatened, and in (20), the “threatenee” is the sun:

- (19) **Afrik-e groz-it isčeznovenie kofejn-yx**
Africa-DAT threaten-PRS.3SG extinction coffee-GEN.PL
derev'-ev.
tree-GEN.PL

‘Africa is threatened by the extinction of coffee-trees.’

¹⁵ Notice that we classify words according to their literal meanings. Thus, *Afrika* in example (19) is classified as “non-person”, even if it arguably may refer metonymically to the people in Africa.

- (20) **Groz-it-sja** napolz-ti na **solnc-e** oblako [...].
threaten-PRS.3SG-REFL cover-INF on sun-ACC cloud
‘A cloud threatened to cover the sun.’

In Table 5, we summarize the situation for the four logical combinations of “person” and “non-person”. The first two rows represent situations where the “threatener” is a person, while the two rows at the bottom involve situations where the “threatener” is not a person.¹⁶

Table 5. Persons and non-persons as arguments

	<i>grozit'</i>	<i>grozit'sja</i>
PersonToPerson	114	210
PersonToNon-Person	0	1
Non-PersonToPerson	83	8
Non-PersonToNon-Person	53	4

The following observations can be made. First, we see that the prototypical threat (PersonToPerson) represents the most frequent option for both verbs. Second, the PersonToNon-Person is marginal. Third, the table shows that *grozit'* is well attested with a non-person as the “threatener”, while this is not the case for *grozit'sja*. In other words, while *grozit'* is relatively evenly distributed between persons and non-persons as the “threatener”, *grozit'sja* strongly prefer persons as arguments. Figure 3 visualizes the difference between persons and non-persons as the “threatener”. The observed difference is statistically significant and has a large effect size.¹⁷

In Section 1 we mentioned that verbs for ‘threaten’ have received considerable attention in studies of grammaticalization since in many European languages the relevant verbs have undergone grammaticalization from the prototypical scenario where one person threatens another person to more abstract meanings involving non-persons as arguments. The Russian data in Table 5 show a similar picture for Russian since non-persons are

¹⁶ Notice that the “threatenees” is not always explicitly marked (e.g., as a grammatical object) in the example sentences. In such cases, we have identified the “threatenees” on the basis of the wider context. The “threatenees” can be realized as noun phrases in different cases, as illustrated in (19) and (20).

¹⁷ We compared examples with Person vs. Non-Person as the “threatener”. Pearson’s Chi-squared test with Yates’s continuity correction (X-squared = 128.87, df = 1) returned a *p*-value < 2.2e-16. Cramer’s V-value is 0.5, indicating a large effect size.

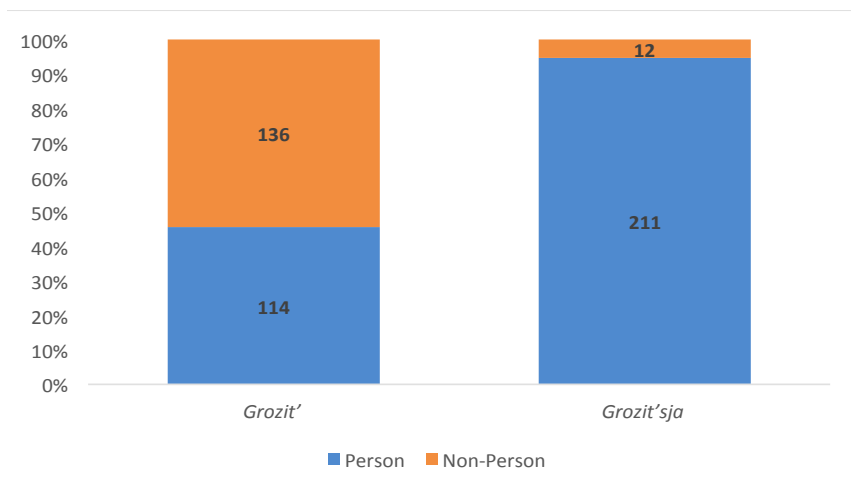


Figure 3. Person vs. Non-Person as “threatener”
(numbers on the bars in the diagram are raw numbers)

widespread. At the same time, Russian is different from the languages for which Heine and Miyashita (2007, 2008) provide detailed analyses because Russian has more than one morphologically related verb for ‘threaten’. Table 5 suggests that it is mainly *grozit'* that follows the path of grammaticalization known from other European languages, whereas *grozit'sja* specializes on the prototypical situation where threats are relations between two persons.

A diachronic analysis lends further support to this conclusion. When we consider the development over time, we see that for *grozit'* the proportion of the PersonToPerson category has decreased over time. In the first half of the 19th century, about 60% of the examples with *grozit'* were of the PersonToPerson type, whereas in the beginning of the 21st century, the corresponding number had decreased to approximately 20%. This difference is statistically significant with a moderate effect size.¹⁸ For *grozit'sja*, on the other hand, the proportion of examples of the PersonToPerson type has been stably high over time. As shown in Table 5 and Figure 3, there are very few examples of the Non-Person type with *grozit'sja*, and this has not changed over time. The historical development is shown in Figure 4 and Table 6 on the following page.

¹⁸ We compared the numbers for *grozit'* in the first half of the 19th century and in the beginning of the 21st century. Pearson’s Chi-squared test with Yates’s continuity correction (X^2 -squared = 16.552, df = 1) returned a p -value = 4.733e-05. Cramer’s V -value was calculated to 0.4.

To summarize, our investigation shows that *grozit'sja* typically describes a relationship between two persons, whereas *grozit'* displays a more varied constructional profile. This difference has increased over time.

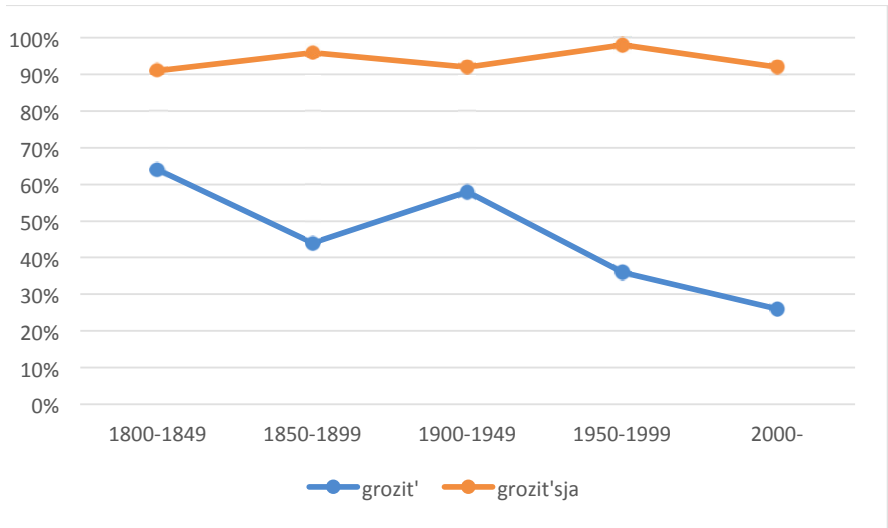


Figure 4. The proportion of the PersonToPerson category over time for *grozit'* and *grozit'sja* (per cent).

Table 6. The proportion of the PersonToPerson category over time for *grozit'* and *grozit'sja* (raw numbers and per cent)¹⁹

	1800–1849	1850–1899	1900–1949	1950–1999	2000–
<i>grozit'</i>	32 (64%)	22 (44%)	29 (58%)	18 (36%)	13 (26%)
<i>grozit'sja</i>	21 (91%)	48 (96%)	46 (92%)	49 (98%)	46 (92%)

6. Verbal vs. Gestural Threats

Having focused on the arguments of the verbs, we now turn to the verbs themselves. In particular, we show that *grozit'sja* tends to involve verbal threats, whereas *grozit'* is more versatile.

Threats can be conveyed by means of words or by a physical gesture, a distinction that has been considered relevant for the choice between *grozit'* and

¹⁹ For each cell in the table, the total is 50 examples. The only exception is the period 1800–1849 for *grozit'sja*, where we have only 23 examples in our database.

In order to test the relevance of verbal vs. gestural threats against corpus data, we annotated our database for three broad categories. “Gestural threats” involve examples where a gesture expressing the threat is explicitly mentioned in the context. The threatener can either use a body part or an object such as a weapon:

- Our category “Verbal threat” covers examples where the threat is conveyed by words and there is no evidence from the context that the threatening words are accompanied by a gesture:

- ²⁰ Notice that the “threatenee” is represented as a prepositional phrase with *na* ‘on’. While the most common pattern is for the “threatenee” to be encoded as a noun phrase in the dative, we have six examples in our database with *na*. The most recent dated example in the Russian National Corpus is from 1937, which suggests that this pattern is somewhat archaic.

The situation is summarized in Table 7, where “other” refers to examples that do not belong to the two categories discussed above, such as sentences where the “threatener” is not a person. As shown, verbal threats are more frequently attested than gestural threats. This holds for both verbs, although the tendency is stronger for *grozit’sja* than for *grozit’*. Contrary to what Gerritsen (1990) proposed, gestural threats are attested for *grozit’sja*, which in our database combines with body parts (e.g., *kulak* ‘fist’, *palec* ‘finger’) and objects (e.g., *skalka* ‘rolling pin’, *arapnik* ‘whip’, *palaš* ‘sword’). Israeli’s suggestion that *grozit’* is the preferred choice for non-verbal threats is supported by our data. For gestural threats, the proportion of examples in our dataset is almost twice as large for *grozit’* compared to *grozit’sja*. The difference is even larger in the category “other”, where we have more than ten times as many examples for *grozit’* as for *grozit’sja*. The large proportion of “other” threats for *grozit’* is related to the fact that *grozit’* often involves non-personal “threateners”, as shown in the previous section.

Table 7. The distribution of verbal and gestural threats for *grozit’* and *grozit’sja*

	<i>grozit’</i>	<i>grozit’sja</i>
Gestural threat	28	15
Verbal threat	81	195
Other	141	13

To summarize, our investigation of data from the Russian National Corpus supports the idea that the distinction between verbal and gestural threats is relevant for *grozit’* and *grozit’sja*. While *grozit’* is frequently used for verbal, gestural, and other threats, for *grozit’sja* verbal threats are the dominant type.

7. Consequences: Serious or Not?

The consequences of a threat may vary from very serious to not serious at all. This has been argued to be relevant for the choice between *grozit’* and *grozit’sja*. In the following, we present weak evidence that *grozit’* may be more compatible with serious consequences, but at the same time, we show that it is difficult to test this hypothesis in a rigorous way.

Commenting on the difference between *grozit’* and *grozit’sja* and similar verb pairs, Israeli (1997: 107) argues that “the non-*sja* verb means an action that has impact”. This is an interesting observation that deserves discussion, although it is far from straightforward to test this hypothesis against corpus

data, since “an action that has impact” can be subjected to various interpretations. However, it seems that we would expect *grozit'* to involve more serious consequences than *grozit'sja*. The question is: what counts as “serious consequences”? Since we are dealing with a scalar phenomenon which is difficult to quantify, it is not easy to avoid subjectivity completely. In order to reduce the level of subjectivity, we decided to focus on the end points of the scale. We divided the examples into three broad categories: “very serious”, “intermediate”, and “non-serious”. In the “very serious” category, we included threats involving death and complete destruction, which are events where the serious effect on the “threatenee” is uncontroversial:

- (25) Naruši-vš-emu zapret-Ø **grozi-l-o**
 break-PST.ACT.PTCP-DAT.M prohibition-ACC threaten-PST-N
 proklat'e i skor-aja smert'-Ø.
 curse and soon-ADJ.F.NOM death-NOM
 ‘Those who did not respect the prohibition were threatened with a curse and an imminent death.’
- (26) Zaduši-t' vsë **grozi-l-Ø-sja**, a potom
 strangle-INF all threaten-PST-M-REFL and later
 i zastreli-l-Ø.
 and shoot-PST-M.SG
 ‘He was threatening to strangle her, and then also shot her dead.’

At the other end of the scale, our category “non-serious” comprises ironic contexts where the threat is not seriously meant. In (27), it is clearly not a real threat that a theater would show “unprecedented decorative installations”. Example (28) is about a party where the invited person promises to come. The use of ‘threat’ instead of ‘promise’ is ironic.

- (27) Teatr **groz-it** pokaza-t' Pariž-u
 theater threaten-PRS.3SG show-INF Paris-DAT
 nevedom-ye dekorativn-ye ustanovk-i.
 unprecedented-ACC.PL decorative-ACC.PL installation-ACC.PL
 ‘The theater threatened to show Paris unprecedented decorative installations.’
- (28) V 10.00 **grozi-l-Ø-sja** by-t' kak štyk.
 at 10.00 threaten-PST-M-REFL be-INF as spit
 ‘At 10.00 he threatened to be there guaranteed.’

The “intermediate” category contains all remaining examples, which cover a whole range of more or less serious threats.

It is instructive to leave the intermediate category aside and compare numbers of the extreme categories “very serious” and “non-serious”, which involve the lowest degree of subjectivity in the classification. Table 8 suggests a tendency for “very serious” threats to favor *grozit'* over *grozit'sja*.

Table 8. Degree of seriousness for *grozit'* and *grozit'sja*

	<i>grozit'</i>	<i>grozit'sja</i>
Very serious	47	23
Intermediate	188	179
Non-serious	15	21

The differences between “very serious” and “non-serious” are statistically significant with a small, but reportable, effect size.²¹ However, we are not dealing with large numbers, and as mentioned, the assessment of the degree of seriousness is to some extent a subjective matter. It is furthermore difficult to control for the interaction with other factors. We conclude that more research is needed in order to better understand the relevance of serious vs. non-serious threats.

8. *Grozit'* vs. *grozit'sja* and the Meaning and Functions of *sja*

What do our findings tell us about the meaning and functions of *sja*? We will argue that *sja* changes the argument structure and the meaning of the verb in a way that relates *grozit'sja* to the middle voice.

Table 9 summarizes our findings. Recall from earlier sections that the results we report are statistical tendencies, rather than categorical rules. Our findings nevertheless show that *grozit'* and *grozit'sja* are semantically and syntactically distinct, although they display overlapping properties. Corpus data therefore clearly represent a valuable resource for the study of *sja*, and a “local approach” studying individual verbs in detail has the potential to sharpen our understanding of *sja*.

²¹ We compared the numbers for “very serious” and “non-serious” threats. Pearson’s Chi-squared test with Yates’s continuity correction (X-squared = 5.3492, df = 1) returned a *p*-value = 0.02. Cramer’s V-value was calculated to 0.2.

Table 9. Overview of findings: Differences between *grozit'* and *grozit'sja* as presented in sections 3 through 7

Topic	<i>grozit'</i>	<i>grozit'sja</i>	Section
Arguments—collocations	nouns	verbs	3
Arguments—syntax	NomVDat	NomVInf	4
Arguments—semantics	person and non-person	person	5
Arguments over time	person decreases	person stays high	5
Situation—type of threat	non-verbal	verbal	6
Consequences	more serious (?)	less serious (?)	7

Does *sja* have an impact on the argument structure of the verb? We first consider the subject. In examples where *sja* serves as a middle (reflexive) marker, a human, or at least animate, subject is required, since such sentences typically involve a human being carrying out a controlled action directed towards oneself. Good examples are “grooming verbs” such as *myt'sja* ‘wash (oneself)’ and *brit'sja* ‘shave (oneself)’. The requirement of a human subject suggests that *grozit'sja* is closely related to examples where *sja* is a middle (reflexive) marker.

With regard to objects, *sja* typically entails decreased transitivity since verbs with *sja* normally do not combine with accusative objects.²² We see subtle effects of *sja* as a “detransitivizer” in many verbs. A case in point is the verb pair *brosat'*–*brosat'sja* ‘throw’, where *brosat'* is transitive and takes an object in the accusative, whereas *brosat'sja* combines with a complement in the instrumental case (Goto and Say 2009: 200; see also Jakobson 1984: 79–80):

²² A small group of verbs like *bojat'sja* ‘fear’ represent an exception to the general rule that *sja* precludes objects in the accusative. For a detailed analysis, the reader is referred to Nessel and Kuznetsova 2015a, 2015b. Notice that we follow Næss (2007) and Letučij (2014), who treat transitivity as a scalar phenomenon structured around a prototype with a direct object in the accusative case (see also Hopper and Thompson 1980 and Chvany 1990). Detailed discussion of this issue is beyond the scope of the present study.

- (29) Kogda **broša-eš'** **kamn-i** v vod-u,
 when throw-PRS.2SG stone-ACC.PL in water-ACC.SG
 sled-i za krug-ami, inače tvoë zanjatie
 watch-IMP.2SG after circle-INS.PL otherwise your activity
 bud-et bessmyslenno.
 be-FUT.3SG meaningless
 'When you throw stones into water, watch the circles, otherwise your activity becomes meaningless.'
- (30) Živu-šč-ij v stekljann-om dom-e
 live-PRS.ACT.PTCP-NOM.M in glass-LOC house-LOC
 ne dolžen-Ø **broša-t'-sja** **kamn-jami.**
 not shall-SG.M throw-INF-REFL stone-INS.PL
 'A person who lives in a glass house should not throw stones.'

Other examples where the accusative object of the non-*sja* verb is demoted to a complement in the instrumental case include *zadavat'sja voprosom* 'ask oneself a question', which corresponds to the synonymous *zadavat' sebe vopros*, with a reflexive pronoun in the dative and a direct object in the accusative (Goto and Say 2009: 194).

A more radical effect of *sja* as a detransitivizer is found in anticausatives such as *slomat'sja* 'break down' and autocausatives like *podnimat'sja* 'get up' (Goto and Say 2009: 194–95). These verbs do not take an object at all, as opposed to the corresponding transitive verbs *slomat'* 'break (something)' and *podnimat'* 'lift (something)' without *sja*, which combine with direct objects in the accusative.

Where does *grozit'sja* place itself in this picture? Does *sja* serve as a "detransitivizer" involving object demotion? Providing a principled answer is not straightforward, since *grozit'* is not a transitive verb with an accusative object. As we have shown, *grozit'* typically combines with a dative and/or instrumental complement, while *grozit'sja* shows an affinity to infinitive complements. Which of these argument structures are most closely related to the transitive prototype with an accusative object? A possible criterion is the ability to undergo passivization. A sentence with a complement that can become the subject of a passive sentence is arguably closer to a prototypical transitive sentence than a sentence where passivization is impossible. Letučij (2014) observes that, in general, infinitive complements have fewer restrictions when it comes to passivization than do nominal complements in other cases than the accusative. If we take this observation seriously, we cannot say that *grozit'sja* is further removed from a prototypical transitive sentence than *grozit'*. At least, *grozit'sja* does not provide strong evidence for the detransitivizing effect of *sja*.

Now that we have considered the arguments of the verbs, we must explore the effect of *sja* on the meaning of the verb itself. We have shown that *grozit'sja* tends to involve verbal threats, possibly with less serious consequences than *grozit'*. It is not straightforward to see a connection to *sja* in other verbs. This, on the other hand, may not come as a big surprise, since Goto and Say (2009: 188) observe that individual semantic idiosyncrasies are quite widespread among verbs with *sja*. In this respect, *grozit'* and *grozit'sja* show similarities with many verb pairs, such as *rešit'-rešit'sja* 'decide'. For instance, both *grozit'sja* and *rešit'sja* often combine with an infinitive complement that has a coreferential subject with the main verb. However, detailed comparison with such verb pairs is beyond the scope of the present study.

To summarize, our analysis shows that *grozit'sja* prefers persons as subject, a feature that relates the verb to verbs where *sja* is a middle voice marker. At the same time, our analysis illustrates the ability of *sja* to change the argument structure and meaning of a verb in somewhat idiosyncratic and unpredictable ways, to some extent dependent on the meaning of the base verb. In other words, simple labels like "middle voice" and "passive" are not sufficient for an adequate analysis. We need detailed studies of individual verbs to arrive at a deeper understanding of *sja*. In short, we need to adopt a "local approach" to verbs with *sja*.

9. Concluding Remarks

In this study, we have provided a thorough analysis of *grozit'* and *grozit'sja*, using corpus data and methods of contemporary corpus linguistics. By way of conclusion, we would like to emphasize the following points. First, we have demonstrated that we are not dealing with complete synonyms because the two verbs under scrutiny differ both syntactically and semantically. We have seen that *grozit'sja* is more likely to combine with human subjects and infinitive complements than is *grozit'*. Moreover, *grozit'sja* tends to involve verbal threats, while *grozit'* is often used about gestural threats. It is furthermore possible that *grozit'sja* implies less serious consequences than does *grozit'*, although we observe that it is difficult to test this hypothesis in a rigorous way.

A second finding concerns *sja*—a descriptively and theoretically challenging morpheme in Russian. Our study testifies to the value of a "local approach" that considers individual verbs in detail. While categories like "middle voice" and "passive" are useful in the analysis of *sja*, we also need detailed analyses of individual verbs in order to pinpoint all the idiosyncratic properties of verbs with *sja*.

Third, our analysis has shown that Russian offers a welcome addition to the theoretical literature on the concept of "threat". In particular, Russian is interesting because it has more than one morphologically related verb for 'threaten'. We have seen that Russian behaves like other European languages

insofar as we find examples of abstract uses that deviate from the prototypical situation where one person promises to do harm to another person. However, our analysis shows that this primarily concerns *grozit'*, while *grozit'sja* is mostly used about prototypical threats.

A final point concerns methodology. Our analysis has illustrated the value of exploring semantic vectors and collocations in large internet corpora like the Araneum Russicum Russicum Maius. However, while these methods can give useful results, they can benefit from being supplemented with detailed analysis of concrete examples, preferably culled from curated corpora like the Russian National Corpus.

Although our analysis suffices to show that *grozit'* and *grozit'sja* are syntactically and semantically distinct, a more detailed analysis of a larger number of examples may shed more light on the differences between the two verbs—and on the meaning and functions of *sja*. In particular, a detailed diachronic analysis of the two verbs would contribute relevant insights, as would a comparison to other Russian verbs for 'threaten', such as imperfective *ugrožat'* and perfective *pogrozit'* and *prigrozit'*. However, these and other issues are beyond the scope of the present study and must be left open for future research.

Sources

Araneum Russicum Russicum Maius. (2015) Russian web corpus. Available at: <http://unesco.uniba.sk/>.

Nacional'nyj korpus russkogo jazyka. (2003–2023) Available at: <https://ruscorpora.ru/>.

References

- Chvany, Catherine V. (1990) "A continuum of lexical transitivity: Slightly-transitive verbs". Martina Björklund, Helena Lundberg, and Janina Orlov, eds. *Carina Amicorum*. Åbo: Åbo Academy Press, 51–63.
- Cornillie, Bert. (2004) "The shift from lexical to subjective readings of Spanish *prometer* 'to promise' and *amenazar* 'to threaten': A corpus-based account". *Pragmatics* 14(1): 1–30.
- Enger, Hans-Olav and Tore Nessel. (1998) The value of cognitive grammar in typological studies: The case of Norwegian and Russian passive, middle, and reflexive". *Nordic journal of linguistics* 22: 27–60.
- Firth, John R. (1957) "A synopsis of linguistic theory 1930–1955". John R. Firth, ed. *Studies in linguistic analysis*. Oxford: Blackwell, 1–32.
- Geniušienė, Emma. (1987) *The typology of reflexives*. Berlin/New York/Amsterdam: Mouton de Gruyter.

- Gerritsen, Nelleke. (1990) *Russian reflexive verbs: In search of unity in diversity*. Amsterdam/Atlanta: Rodopi.
- Glovinskaja, Marija Ja. (2004a) "Grozit'". Ju. D. Apresjan et al., eds. *Novyj ob'jasnitel'nyj slovar' sinonimov russkogo jazyka*. Moscow: Jazyki slavjanskoj kul'tury, 240–43.
- . (2004b) "Ugrožat'". Ju. D. Apresjan et al., eds. *Novyj ob'jasnitel'nyj slovar' russkogo jazyka*. Moscow: Jazyki slavjanskoj kul'tury, 1190–94.
- Goto, Ksenija V. and Sergey S. Say. (2009) "Častotnye xarakteristiki klassov russkix reflektivnyx glagolov". Ksenija L. Kiseleva, V. A. Plungian, E. V. Raxilina, and S. G. Tatevosov, eds. *Korpusnye issledovanija po russkoj grammatike*. Moscow: Probel-2000, 184–223.
- Harris, Zellig S. (1954) "Distributional structure". *Word* 10(2–3): 146–62.
- Heine, Bernd and Hiroyuki Miyashita. (2007) "The structure of a functional category: German *drohen*". *CogniTextes* 1: 1–23.
- . (2008) "Accounting for a functional category: German *drohen* 'to threaten'". *Language sciences* 30: 53–101.
- Hopper, Paul J. and Sandra A. Thompson. (1980) "Transitivity in grammar and discourse". *Language* 56(2): 251–99.
- Israeli, Alina. (1997) *Semantics and pragmatics of the "reflexive" verbs in Russian*. Munich: Otto Sagner.
- Jakobson, Roman O. (1984) "Contribution to the general theory of case: General meanings of the Russian cases". Linda R. Waugh and Morris Halle, eds. *Russian and Slavic grammar: Studies 1931–1981*. Berlin/New York/Amsterdam: Mouton, 59–104.
- Joos, Martin. (1950) "Description of language design". *Journal of the Acoustic Society of America* 22(6): 701–08.
- Jurafsky, Daniel and James H. Martin. (2021) *Speech and language processing*. See esp. chap. 6, "Vector Semantics and Embeddings". Draft of 3 February 2024. Available at: <https://web.stanford.edu/~jurafsky/slp3/>.
- Kemmer, Suzanne. (1993) *The middle voice*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Knjazev, Jurij P. (2007) *Grammatičeskaja semantika: Russkij jazyk v tipologičkoj perspektive*. Moscow: Jazyki Slavjanskix Kultur.
- Kyröläinen, Aki-Juhani. (2013) *Reflexive space: A constructionist model of the Russian reflexive marker*. Ph.D. dissertation, University of Turku.
- Langacker, Ronald W. (1991) *Concept, image, and symbol*. Berlin: Mouton de Gruyter.
- . (1999) *Grammar and conceptualization*. Berlin: Mouton de Gruyter.
- Letučij, Aleksandr B. (2014) "Perexodnost'". *Russkaja korpusnaja grammatika*. Available at: <http://rusgram.ru/%D0%9F%D0%B5%D1%80%D0%B5%D1%85%D0%BE%D0%B4%D0%BD%D0%BE%D1%81%D1%82%D1%8C>.

- Makarova, Anastasia and Tore Nessel. (2024) "Replication Data for: Threatening in Russian with or without *-sja*: *grozit'* vs. *grozit'sja*". Available at: <https://doi.org/10.18710/J3UCVC>. DataverseNO, V1.
- Narrog, Heiko and Bernd Heine. (2021) *Grammaticalization*. Oxford: Oxford University Press.
- Nessel, Tore and Julia Kuznetsova. (2015a) "Constructions and language change: From genitive to accusative objects in Russian". *Diachronica* 32(3): 365–96.
- . (2015b) "In which case are Russians afraid? *Bojat'sja* with genitive and accusative objects". *Journal of Slavic linguistics* 23(2): 255–83.
- Næss, Åshild. (2007) *Prototypical transitivity*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Peškovskij, Aleksandr M. (1956) *Russkij sintaksis v naučnom osveščanii*. 7th edition. Moscow: Gosudarstvennoe učebno-pedagogičeskoe isdatel'stvo ministerstva prosvěščenija RSFSR.
- Verhagen, Arie. (1995) "Subjectification, syntax, and communication". Dieter Stein and Susan Wright, eds. *Subjectivity and subjectivisation: Linguistic perspectives*. Cambridge, UK: Cambridge University Press.

Tore Nessel
Department of Language and Culture
UiT The Arctic University of Norway
Tromsø, Norway
tore.nessel@uit.no

Anastasia Makarova
Department of Modern Languages
Uppsala University
Uppsala, Sweden
anastasia.makarova@moderna.uu.se

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.

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 corpus⁶ (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.

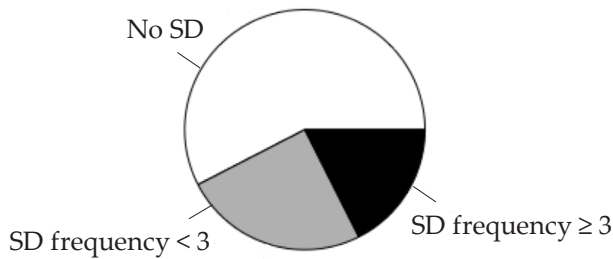


Figure 1. The distribution of SD frequency across speakers (per 1,000 tokens)

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

	<i>t</i> -stem	<i>v</i> -stem	<i>n</i> -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 (χ -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
<i>t</i> -stem	15	1000	90	1105
<i>v</i> -stem	154	8	5	167
<i>n</i> -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

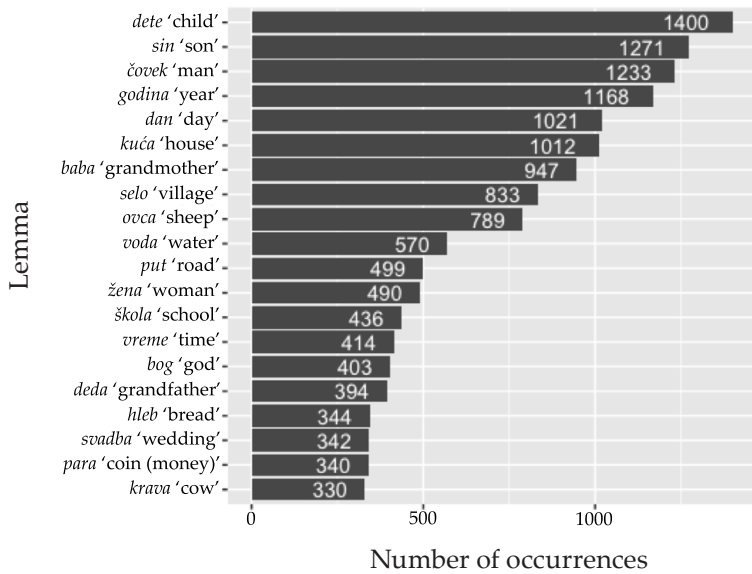
	Max (abs freq)	Mean (abs freq)	SD (abs freq)	CI LL	CI UL
Bare noun	1,400	5.48	33.35	97.50%	97.90%
Noun + SD	27	0.07	0.77	2.07%	2.52%

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.

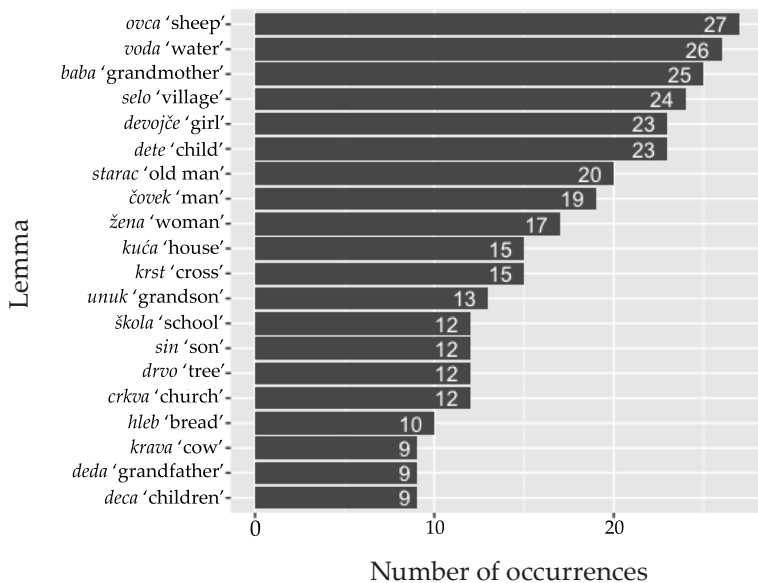
¹⁴ Function “glm()” was used from the R package Stats (R Core Team 2022).

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	<i>t</i>-value	<i>p</i>-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 ($N_{\text{quant}} = 9$), 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 ($N_{\text{ordnum}} = 5$), and in that case, the SD attaches to the numeral functioning as an adjectival modifier, (7b). There are four occurrences of quantifiers like *oba/objica* ‘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. objica-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 ($N_{\text{adj}} = 13$), 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 člana v made-donskom 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 rečnik* [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 knjižoven 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. *Bългарistični proučvaniji* [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 literaturni 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 *тъ und seine Rolle bei der Grammatikalisierung von Definitheit im Slavischen" [The pronoun *тъ 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 sãvremennija bãlgarski knižoven ezik* [Grammar of the contemporary Bulgarian literary language]. Vol. 2. Sofia: Bãlgarskata akademija na naukite.
- Šimko, Ivan. (2020) "Definiteness markers in the *Life of St. Petka*". *Zeitschrift für Slavistik* 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

Reviews

Senahid Halilović (chief of the project), Mehmed Kardaš, Amela Ljevo-Ovčina, and Emira Mešanović-Meša. *Bosanskohercegovački lingvistički atlas I: Fonetika* [Bosnian-Hercegovinian linguistic atlas I: Phonetics]. Sarajevo: Slavistički komitet, 2020. 428 pp. ISBN 978-9958-648-28-1. Available at: https://slavisticki-komitet.ba/Bosanskohercegovački_lingvisticki_atlas.pdf.

Reviewed by Ronelle Alexander

Dialectology, in the broadest sense, studies language variation. Although the term “dialect” can have several referents, it most often denotes traditional rural speech, as unaffected as possible by interference from the standard language or other contact elements. Linguists normally record this speech in situ, either by immersing themselves in the local speech through extensive residence (and then writing a full grammar of this local speech system), or by visiting a number of different areas and then comparing the results along different parameters. The most consistent and reliable way to undertake the latter is first to draw up a network of locations that is presumed to give a relatively thorough geographical coverage, then to make a detailed outline of the desired data (usually as a list of words which contain the desired phonemes or morphemes), and finally to record (to the extent possible) comparable data in each locale. The fact that the data outline is commonly called a questionnaire conjures up the unfortunate image of interrogation. Although there is sometimes no other way to elicit the desired item than by asking a direct question, much of the work can be done by simply directing conversation towards topics in which the desired forms might reasonably be uttered.

The resulting data can be mined for studies analyzing specific questions, but such data are most frequently presented to the public en masse in the form of a dialect atlas. This is a comprehensive volume containing a large number of maps, usually with some sort of commentary. Each map is devoted to one of the data items, and each of the relevant responses is displayed on the map at the geographical point where it was recorded. Although it takes a great amount of work to collate the data and construct these maps, the result is highly satisfying, allowing the reader to visualize dialectal differentiation in a vivid and direct manner.

In Slavic countries, the socialist period was particularly productive in terms of dialect atlases, largely due to the extensive support provided by socialist governments (for the correspondingly extensive amount of work such an enterprise requires); and this resulted in the publication of atlases for most of the Slavic languages, though of various formats and with a varying extent of coverage. A notable lacuna was the language formerly called Serbo-Croatian. Although Serbo-Croatian dialectology was a vibrant, active, and highly respected discipline throughout Yugoslavia's lifetime (and especially during the socialist period), conditions were not conducive to the production of an atlas which would cover the full range of what is now called BC(M)S. The difficulties were both practical and theoretical: not only was scholarly collaboration across the country hampered by the lack of inter-republic coordination and the increasing decentralization of the federation, but dialectologists themselves were strongly influenced by the interwar "deans" of dialectology, Aleksandar Belić in Serbia and Stjepan Ivšić in Croatia, both of whom disdained questionnaire work as excessively mechanical and artificial, and insisted their students instead gather data by the immersion method.

There may be (or may have been) some justification for this point of view, but most dialectologists are sufficiently adept at fieldwork to make the data collection more natural than mechanical, and they also develop the intuition needed to distinguish "authentic" responses from "artificial" ones. However, it is indeed important to note that both methods are necessary to obtain maximal information about dialects. The immersion method, which allows the writing of a comprehensive grammar of the dialect, not only provides cultural context, but also descriptions of syntax (and indeed of all linguistic data beyond the level of what can be illustrated by single-word examples). Still, when it comes to the nuts and bolts of language (phonology, morphology, and the lexicon), there is nothing more impressive than a solidly constructed dialect atlas, and nothing quite so satisfying as the visual and intellectual pleasure of a well-drawn dialect map.

It is a joy, then, to hear of the publication of volume 1 (*Fonetika*) of the *Bosanskohercegovački lingvistički atlas*, available online in PDF format. The data, representing 230 villages, are drawn largely from material gathered between 1975 and 1986, some published then and the rest retained in archives. Six more villages were investigated in 2016–17, four for the first time and two as a follow-up. In his foreword, the director of the project, Senahid Halilović, acknowledges the "significant changes in the dialectal situation due to population movements" occasioned by the wars of succession but asserts that the two sets of data are sufficiently comparable to allow for the preparation of a linguistic atlas. It is highly doubtful that the very precise and close-grained variation displayed on the maps now being presented, maps admittedly drawn from pre-war data, is a truthful representation of the current post-migration situation (nor does Halilović make such a claim; all he says is that the

small amount of new information gathered in 2016–17 is “comparable”). What is important is that the data are consistent.

The metadata preceding the actual maps is thorough. First, there are three different listings of the sites investigated (each ordered differently), followed by two lists of personnel responsible for the data: one list identifies authorship of the data (who did the actual investigation and recording of data), and the other identifies authorship of individual maps (who was responsible for collating the data and creating the map). Second, there is a full description of the transcription system used, which takes care to note that it is consistent with the transcription system used in those European atlases in which Bosnian material is represented. Finally, there is a full list of the lexemes on the questionnaire (with translation into English, French, German, and Russian). There is also an extensive appendix, a 32-page alphabetical list containing every single form listed on any map (in phonetic transcription) with reference to the map which displays it, and English, French, German, and Russian summaries of Halilović's foreword.

It is the maps themselves, of course, which constitute the core of any atlas. These are introduced by a map delineating boundaries of the four basic dialect groups (East Bosnian, East Herzegovinian, Western, and Posavian). The fact that the latter group includes only two of the 230 villages investigated is a sad reminder of the costs of the breakup, as the major part of this historically very important dialect group lies to the north of the Sava, in the Slavonian section of Croatia; consequently, it is now under the purview of Croatian dialectologists and not “available” to those in Bosnia.

However, these Bosnian dialectologists have done a masterful job with their own material. The introductory section includes two more maps, on which are plotted all 230 of the investigated points, with the numbers and abbreviated names of each rendered in one of three colors. This same color scheme, with green representing Bosniak villages (100 in all), red representing Serbian villages (80 in all), and blue representing Croatian villages (50 in all), is carried through on all the maps (and does appear, by the way, to represent the pre-war distribution of ethnicity). Each of the 181 maps is devoted to a particular questionnaire item, with 63 of them devoted to vocalic phenomena and 118 to consonantal phenomena. The presentation of each map covers two pages: the first gives a list of all the responses to the relevant question, followed by linguistic and etymological commentary, and the second contains the map itself. A further feature of each map page is the presence of four pie charts to the left, intended to give a bird's-eye view of the distribution of the several reflexes. The first depicts the overall distribution, and each of the subsequent three depicts the distribution within the set of villages ascribed to each of the three ethnicities.

At first glance, it may seem excessive to pay such detailed attention to ethnic differentiation. Such differentiation, however, is a fact of life in post-Dayton

Bosnia. Furthermore, even in the prewar period, when it was assumed that everyone spoke something akin to Serbo-Croatian, it was well known that such differences existed in dialectal speech. Of course, these differences were at that point marked as characteristic of the speech of Muslims, Catholics, or Orthodox, and not (as they are herein) as characteristic of the speech of Bosniaks, Croats, or Serbs, respectively.

The maps themselves are excellently drawn, making a vast amount of information clearly accessible to the reader. One feature is curious, however: while all maps define the several types of responses to the question being depicted and mark each village on the map with the relevant symbol, only some provide a further level of graphic information. This further level consists in the colored shading (usually yellow) of the area of one particular response, which functions to set this area clearly off from the remainder (a very few instances use two or even three different colors). There is no reasoning given, nor any that could be intuited, as to why only some maps are constructed this way, so it remains a head-scratcher. All the maps are valuable, though, with or without the added value of what amounts to the drawing of an isogloss.

For instance, the map on pp. 164–65 is intended to show which localities preserve the palatal *L* in *ulje* ‘oil’ and which turn it into *j* (*uje*); but it also lets us see that a large proportion of all places in Bosnia say neither one but use a different vocabulary item, *zejtin* (which came in through Turkish). The map on pp. 202–03 gives us data to test the old stereotype about the word for ‘coffee’, which is that Muslim Bosniaks say *kahva*, Croats say *kava*, and Serbs *kafa*. In fact, there is a significant correlation between ethnic identification and choice of ‘coffee’ words, but it is far from being 100%.

In sum, this first atlas to appear within the larger “central South Slavic” region is a very welcome addition to scholarship, and the compilers are to be congratulated. One awaits further volumes with great anticipation.

Ronelle Alexander
Department of Slavic Languages and Literatures
University of California, Berkeley
Berkeley, CA, USA
ralex@berkeley.edu

Thomas Rosén. *Russian in the 1740s*. Boston, MA: Academic Studies Press, 2022. xiv + 198 pp. ISBN 9781644694145 (hardback); 9781644694979 (paperback); 9781644694169 (ePub); 9781644698303 (Open Access).

Reviewed by John Dunn

The Russian language in the 18th century can be compared to a sausage: we know pretty well what ingredients are used and we have an exact knowledge of what the final product is like, but what is less well understood is the bit in between. There is still much to discover about the processes by which the raw ingredients—in this case, the various forms of Russian and Church Slavonic that co-existed in Muscovite Russia at the beginning of the 18th century—are converted into this final product, the linguistic variety recognizable as something close to Modern Russian that emerges just over a century later. Dr. Rosén seeks to expand our knowledge and understanding of some of these processes by concentrating, as the title indicates, on a specific decade, the 1740s.

The choice of the 1740s is in part serendipitous, in that it was initially prompted by the accidental discovery of some Russian letters dating from that decade in the Swedish National Archives, but it is also a decade that is relatively unstudied, falling, as it does, between the linguistic upheavals of the Petrine period and Lomonosov's interventions of the 1750s. The lack of previous study might presuppose two potential outcomes: either the emergence of a quantity of previously unconsidered material offering useful insights, or else the sad discovery that there is not very much material available to study; in a curious and unexpected way, this book suggests that, with regard to this particular case, both outcomes might be possible.

The book is divided into nine chapters. After two introductory chapters, one giving general background information and the second discussing previous research, chapter 3 is devoted to an examination of the socio-linguistic situation in the Russia of the 1740s. Chapters 4–6 contain preliminary material relating to the texts which are to be analyzed, dealing with such matters as available sources, methodological issues, and what the author calls the “Situational Analysis of Registers”; this rather cryptic term refers to the participants and the relationships between them, the communicative purposes of the documents, the topics discussed, and other related matters. The linguistic analysis itself is the subject of chapter 7, and this is followed by a rather inconsequential chapter entitled “Functional Analysis” and a brief final chapter

giving general conclusions. What this outline may well suggest is a certain imbalance between introduction and presentation on one hand, and analysis on the other. The most important sections of the book are chapters 3 and 7, and it is on these that the greatest attention will be focused in the remainder of this review.

Chapter 3 is essentially divided into two parts: the first is concerned with education and literacy, while the second deals with what the author calls language management. On education, we are told rather more about structures than about content, which is perhaps inevitable given the nature of the information available. On literacy, the available information is even less helpful; though the author is able to produce interesting material relating to two Russian regiments, he is for the most part reduced to conjecture, based to a large extent on evidence relating to periods other than that which he is studying. The section on language management is likewise unenlightening. We are given a great deal of information about the structures and personnel of the Academy of Sciences and its *Rossijskoe sobranie* (Russian Conference); we are even given Tredjakovskij's contract of employment with the Academy in both French and Russian. For all that, however, we learn almost nothing about what these structures did in terms of language management, and it thus becomes hard to avoid the conclusion that in the 1740s there was little or no activity relating to language management, or if there was, it was carried out in places other than the Academy of Sciences.

It is true that, from time to time, the author suggests lines of enquiry that remain unexplored. In §3.2.2 (p. 42), he notes the probability that a significant part of the population might have been able to read Church Slavonic but not Russian, without, however, considering what this might mean in practice or what the implications might be. In §3.3.2 (p. 51), he mentions but does not analyze or discuss Tredjakovskij's 1748 treatise on orthography (which one might have thought an attempt at language management), and he concludes the chapter by quoting in full two template documents produced by the state administration: one relating to the Imperial title, the other being for a letter of credit. This is potentially a very useful resource, but Dr. Rosén appears more interested in the fact of their existence rather than the details of their linguistic content, which leaves matters rather hanging in the air. Nevertheless, for all these unexplored avenues, the reader may well feel that this chapter arrives perforce at the second of the two outcomes mentioned earlier.

The choice of texts subject to linguistic analysis is inevitably influenced by considerations of availability, including the discoveries made by the author himself. The texts encompass both manuscript and printed documents and include a hand-written receipt; a letter and an official report relating to a naval incident in 1742; a number of diplomatic documents, including royal correspondence, preserved in the Swedish and Danish National Archives; an extract from the *Artikul voinskij* (*Артикул воинский*, the military regulations

of Peter I, first issued in 1715); and an extract from Field Marshal de Lacy's reports from the front during the Russo-Swedish War of 1741–43. It is greatly to Dr. Rosén's credit that he quotes copiously from his selected texts, and where it is appropriate to do so, he places parallel texts in adjacent columns to facilitate comparisons (as with the three editions of the *Artikul voinskij*). Only once does this system break down: he quotes a lengthy extract from the de Lacy text, but then much of the discussion centers round passages from elsewhere in the document, which diminishes the relevance of the extract.

Here, there is a great deal to discover, but it has to be said that the quality of the linguistic analysis is somewhat uneven. The author seems to be mainly interested in orthography and morphology, and while he makes a number of useful observations on the former, on the latter he is too often content merely to note the presence of phenomena, without exploring their significance. Sometimes he does not even manage that; discussing the documents relating to the 1742 incident, he merely notes that "[i]n terms of morphology, the language of the documents contains little that cannot be expected from a text of the 1740s" (p. 125). Particularly disappointing is the lack of attention paid to syntax and vocabulary, both of which one might feel to be particularly important for texts of this nature and from this period. If the occasional syntactic construction is mentioned, vocabulary is almost totally ignored. For example, a letter from Empress Elisaveta Petrovna to the King of Sweden, quoted in full on pp. 130–32, contains a considerable number of striking syntactic and lexical elements, but while Dr. Rosén rightly draws attention to the use of formulae, these specific elements go unnoticed. Similarly, in the discussion of the three editions of the *Artikul voinskij*, the orthographical and morphological differences are conscientiously noted (other aspects of the language remain unaltered), but what escapes the author's attention is that these changes, and especially those introduced in the 1744 edition, provide useful information relating to the processes leading to the standardization of Russian orthography.

The reader may thus come away from this book with the feeling that an opportunity has been missed here and that the author has not fully succeeded in achieving what he set out to do. Perhaps the problem is that in attempting both to explore the sociolinguistic situation of Russia in the 1740s and to offer a detailed linguistic analysis of a number of individual texts, the work ends up falling between two stools, so that neither task is accomplished as well as it might have been. That said, however, the book should not be written off, since all those concerned with the Russian language of this period will find here much that will be of value, above all, perhaps, as a springboard for further research. Moreover, they will certainly have cause to be extremely grateful to Dr. Rosén for generously reproducing so much textual material and for presenting it in such a user-friendly manner.

The book is produced to a very high standard, and both the author and the publishers are to be warmly congratulated on the care they have taken

over the reproduction of texts that must have been extremely difficult to handle. The author's English, while not quite up to native-speaker standard, is certainly more than serviceable, but there is one mystery: how on earth was it possible that throughout the entire process of writing, editing, and publishing this book, nobody noticed that the word "genitive" is consistently misspelled, an error made all the more visible by the presence of a redundant capital letter?

John Dunn
School of Modern Languages and Cultures
University of Glasgow
Glasgow, Scotland
john.dunn@glasgow.ac.uk

Anna-Maria Sonnemann. *Language Families in Contact: The Mutual Impact of Slavic and Romani*. Berlin/Boston: DeGruyter Mouton, 2022. xvii + 247 pp. [Language Contact and Bilingualism, ed. Yaron Matras, volume 24.] ISBN 978-3-11-075604-3.

Reviewed by Victor A. Friedman

For specialists in Slavic linguistics, the study of Slavic contacts with Romani is important for discussions of the kinds of grammatical change that can occur in a non-Slavic system under the influence of various Slavic languages, as well as changes in the lexicon in both directions. This useful work, based on the author's habilitation thesis at the University of Cologne (2021)—which itself expands and revises some of the author's previously published work—provides an excellent survey of all the key issues. Aside from the usual front matter (i–xviii), the book consists of nine chapters (1–170), two appendices (171–207), about 365 references (219–39), a subject index (241–43), and a language index (245–47).

Chapter 1 (1–15) is the Introduction, which gives a brief discussion of the problems of determining the number of Romani speakers (as opposed to those who identify as Romani, itself also problematic due to underreporting connected with social stigma), a brief discussion of Romani multilingualism, the history of research on Slavic-Romani language contact, the Romani Morphosyntactic Database (RMS), which is indeed the single most useful comprehensive source for looking at Romani-Slavic contact, and an overview of the volume. Chapter 2 (16–25) presents the author's arguments for treating Romani as a “family” on the same level as the Slavic “family”. Chapter 3 (26–32) is entitled “Examining contact phenomena through a contemporary lens”. Chapter 4 (33–66) gives an overview of the structural impact of Slavic on Romani. Chapter 5 (67–91) looks at Slavic verbal prefixes in Romani. Chapter 6 (92–113) discusses Slavic lexical borrowings in Romani, and Chapter 7 (114–38) examines Romani borrowings in a selection of Slavic languages (BCMS, Bulgarian, Czech, Polish, Russian, and Ukrainian). Chapter 8 (114–38) is entitled “Writing Romani in ‘Slavic’ alphabets”, and Chapter 9 (165–70) gives the conclusions.

Chapter 1 does a good job of covering the basics necessary for this book. Chapter 2 is somewhat problematic. The author begins by asking: “What is a language family, what is a language, what is a dialect, and where do

Romani and the Slavic languages belong in this framework?" The discussion of the term *family* as having broad and narrow definitions is adequate. Here I would emphasize the definition used by Nichols (1992: 24–25) that is in one of Sonnemann's citations. Nichols uses *family* for the time depth of one of the older branches of Indo-European (2500–4000 years, e.g. Iranian) and *stock* for the deepest reconstructible phylogenetic node, e.g., Indo-European. We can note in passing that this frees up *phylum* for a hypothetical deeper genealogical node for which some evidence can be adduced, even though the signal is weaker than that required for the stock, e.g., (Na-)Dene-Yenisean (Kari and Potter 2010). Sonnemann rightly skates over the problem of whether or not Balto-Slavic is a reconstructible unity, and does not bother to mention the question of whether Common Slavic as we have reconstructed it was a relatively late koine that was superimposed over earlier dialectal diversity, as happened with Ancient Greek, when the Attic-based koine displaced all the ancient dialects except, to some extent, the Doric that is the basis of Tsakonian. Even if reconstructed Common Slavic as we know it is considerably more recent than, e.g., Proto-Iranian, I see no problem using the same term for the genealogical level of Slavic. Balto-Slavic would then be at the level of Indo-Iranian, even if the reconstruction of the latter is more secure than that of the former.

Treating Romani as a "family" on the same level as Slavic, however, is problematic. Basically, the author compares 15 sentences of a dialect from Bulgaria with a dialect from Finland, finds that 50 out of 90 words (56%) are neither identical nor cognate, and concludes that despite a variety of basic phonological, grammatical, and lexical commonalities among Romani dialects, Romani is a "family" like Slavic. She does not pursue this thought, however, to specify which dialects enter into which "languages", nor does she consider the sociolinguistic implications of her claim, nor the terminological problem of what to call Indic or Indo-Iranian if Romani is a "family" and Indo-European is a "stock". For the rest of the book, no attempt is made to assign the various dialects to "languages". Overall, this chapter does not really contribute much.

Chapter 3 is a quick look at some contemporary theories of language contact, of which the author chooses the functional-pragmatic approach exemplified by Matras (2020). This is certainly a suitable approach for the data. Chapter 4 looks at the structural impact of Slavic on Romani and examines the following topics: phonetics and phonology (33–37); nominal morphology (37): inflectional endings and word-formation affixes (37–38), cases (38–41), articles (41–42), comparison (42–43), indefinites (43–45), interrogatives (45), possessives, personal pronouns, and demonstratives (46); verbal morphology (46–47): tense (47–48), the "new infinitive" [*sic*; I see no reason for the scare quotes] (48–49), voice and reflexivity (50–51), modality (51–53), the conditional (53), renarrative and evidentiality (53–54), imperative, optative, and jussive (54); syntax (54): prepositions and conjunctions (55–58), object doubling (58–59), conditional

sentences (59–60), negation (60), word order (60–62), and zero copula (62–63); and a summary (63–66). In this chapter, the author does a very competent job of presenting the most important phenomena for each topic, with useful examples for every point. This chapter represents an excellent overview of the subject and would be suitable for assignment in Slavic, Romani, and contact linguistics courses. A very small quibble with the summary (p. 60): not all the Romani dialects in contact with Macedonian have final devoicing, although some do. (This is actually a topic that requires future research.)

Chapter 5 is an insightful discussion of Slavic verbal prefixes in Romani, based mostly on the RMS database. Sonnemann identifies three regional patterns in prefix borrowing, moving from south to north corresponding to from least to most. Dialects in contact with South Slavic have very limited borrowings of Slavic prefixes on Romani verbs; the numbers are higher for dialects in contact with Czech and Slovak, and highest in dialects in contact with Polish and East Slavic. No Romani dialect, however, actually imports the Slavic perfective/imperfective aspectual system. Rather, the use of prefixes for Aktionsarten is what gets borrowed to a lesser or greater extent as one moves from south to north.

Chapter 6 gives an overview of Slavic lexical borrowings into Romani dialects currently in contact with Slavic, and a more detailed case study of two Polish varieties, Bergitka and Polska, where Bergitka borrows more heavily from Polish than Polska. There is also a brief section on calquing. The first part of the chapter is organized by word class. This section contains several bar charts of lexical meanings showing the number of dialects borrowing a given type of lexical item using English glosses, e.g., for temporal and phasal adverbs, the largest number—over 60—borrow “always” while the smallest number—one or two (the chart is graded by units of 10)—borrow “today”. Other charts are given for local adverbs and for focus particles and intensifiers. There are similar charts for prepositions and for conjunctions in Chapter 4. It would be interesting to compare such data with a larger typological study such as that in Haspelmath and Tadmor (2009 a, b), or Tadmor et al. (2010).

Chapter 7 is entitled “Romani borrowings in diastatic varieties of Slavic”. The qualification “diastatic” is necessary since borrowings from Romani into Slavic, even the most widespread and everyday, are limited to colloquial registers (or their written representations), cryptoglossic, and slang registers of various sorts. The chapter gives a brief survey of sources of such elements in specific argots of Bulgarian, BCMS, Czech, Polish, Russian and Ukrainian, e.g., BCMS *šatrovački*, Bulgarian *Čalgadžiski ezik*, etc. The author then examines Romani elements in these languages’ modern colloquial registers. For BCMS, she gives a bar chart of frequency counts in the Croatian national corpus. An interesting point that emerges is that while *love* (Romani *love* is borrowed into BCMS as *lova*) ‘money’ is the most frequent in BCMS, it is completely absent from Bulgarian. The discussion of Bulgarian *aver* ‘friend, companion’

spends too much time on Leschber's (2002) unconvincing arguments for a Romani origin for this term (Romani *aver* 'other'). The Bulgarian is surely ultimately from Hebrew *kha'aver* 'friend, companion, comrade', and more likely via Judezmo rather than Yiddish, which is proposed as a possible source on 129. The word is also listed in Appendix 1, and although it is set off in square brackets, there is no explanation why this is a very doubtful entry. Frequency bar charts of Romani words are also given for the Czech national corpus and the Russian national corpus. An especially useful feature of this chapter is the comparison of the historical material registered in studies of argots with modern material from national corpora and other contemporary sources.

Chapter 8 addresses the issues of writing Romani in nation-states where the dominant language is Slavic. It represents a good overview of the various issues connected with writing Romani, which, until the twentieth century, was generally written only by linguists rather than by speakers themselves. Chapter 9 summarizes the book's contents, Appendix 1 lists the Polish words in the two Romani dialects discussed in detail in Chapter 6, and Appendix 2 gives a useful list of Romani words in the six Slavic languages considered in Chapter 7, with columns of specifications: one for historically attested argots and the other for modern usage, complete with the forms these various words take in the recipient languages.

There is much that is original in this book in terms of its research. It is a contribution of new knowledge. Slavic linguists interested in language contact will be able to read and use this book with profit.

References

- Haspelmath, Martin and Uri Tadmor, eds. (2009a) *World loanword database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: <https://wold.clld.org/>.
- . (2009b) *Loanwords in the world's languages: A comparative handbook*. Berlin: De Gruyter.
- Kari, James and Ben A. Potter, eds. (2010) *The Dene-Yenisean connection*. Fairbanks: University of Alaska.
- Leschber, Corinna. (2002) "Semantische vorgänge bei lexikalischen übernahmen aus dem Romani in diastratische varietäten des Bulgarischen". [Semantic processes in lexical adoptions from Romani into diastratic varieties of Bulgarian.] *Grazer Linguistische Studien* 58: 57–101.
- Matras, Yaron. (2020) *Language contact*. Cambridge, UK: Cambridge University Press. 2nd ed.
- Nichols, Johanna. (1992) *Linguistic diversity in space and time*. Chicago: University of Chicago.

- RMS = Matras, Yaron and Viktor Elšík. (2001–05, 2008) Romani Morpho-Syntax Database. University of Manchester. Available at: <https://romani.humanities.manchester.ac.uk/rms/>.
- Tadmor, Uri, Martin Haspelmath, and Bradley Taylor. (2010) "Borrowability and the notion of basic vocabulary". *Diachronica* 27(2): 226–46.

Victor A. Friedman
University of Chicago and La Trobe University
Chicago, IL, USA, and Melbourne, Australia
vfriedm@uchicago.edu

Ágoston Pável. (Marc L. Greenberg, ed./trans.) *Prekmurje Slovene Grammar: Avgust Pavel's Vend Nyelvtan (1942)*. Leiden: Brill-Rodopi, 2020. xxvii + 215 pp. ISBN 978-90-04-41911-7.

Reviewed by Grant H. Lundberg

Marc Greenberg's translation and annotation of Avgust Pavel's 1942 *Prekmurje Slovene Grammar* is important for several reasons. First, it is often noted that the Slovene speech territory is a contact zone for the major language families of Europe. It is the place where Slavic, Romance, Germanic, and Hungarian meet. This is true in a more concentrated way for the Prekmurje region of Slovene. In this area, the languages and dialects of Slovene, German, and Hungarian are in intense contact. This is a region where national borders have often shifted, and linguistic minorities have been strongly influenced by the titular languages of the countries they were located in. Greenberg's work makes important linguistic details more widely available. Second, the grammar was written at an important time. Pavel's work captures the state of the dialect region and literary Prekmurje Slovene at a time (1942) before the onset of heavy influence from Standard/Central Slovene, which would soon follow. The grammar preserves evidence of linguistic forms that are no longer observed. This provides valuable comparative material for the study of Slavic dialects and language history. Third, the grammar was translated and annotated by the recognized expert on the phonological history of Slovene and its dialects. He literally wrote the book on the topic. Greenberg's introduction to the translation is a careful summary of all the important topics in the phonological history of the Prekmurje language. Along with the bibliography, the introduction and commentary serve as a foundational resource for the study of Prekmurje Slovene.

Greenberg's introduction to the translation is a brief discussion of the Prekmurje dialect and its early differentiation, along with Pannonian dialects, from other neighboring Slovene and Kajkavian dialects, as well as a sketch of Prekmurje's role in the dialect continuum between West and South Slavic (p. 1). There is an especially interesting review of research on the distribution of variants of several verbal suffixes and lexical items that point to dialect divisions in Slavic before the great migration around 500 AD (3–4). Greenberg also covers the development of the vocalic systems in Prekmurje in the context of South Slavic, as well as the internal differentiation of Prekmurje (4–5). This

is complicated material presented by a specialist who has mastered dialect and historical details as well as the broad Slavic and Indo-European context. Finally, the introduction also contains a brief discussion of the history of the written tradition of Prekmurje Slovene and Pavel's role in the failed Hungarian efforts to standardize the language (2).

Greenberg's annotation within the grammar is interesting and insightful. He conducts it like a conversation with Pavel's material, commenting on and explaining ambiguous statements, sometimes confirming Pavel's observations with his or others' dialect research, sometimes correcting historical claims based on etymological research completed after Pavel's time. I will give several illustrative examples below.

In his discussion of the phonetics and phonology of Prekmurje Slovene, Pavel indicates that expected long vowels are shortened when the syllable is closed by a nasal consonant. Greenberg notes that this is a feature of the Cankova dialect, which Pavel generalizes for the entire grammar, "though there is no evidence that this shortening rule applies to other parts of Prekmurje" (24). Even though it is not the aim of the grammar, on several occasions Pavel points out what he understands to be the derivational or lexical history of modern forms, which Greenberg corrects based on etymological and dialect research: "In the particular examples he adduces, however, the derivations are not formed as he claims" (30). In another example, Pavel points out the origin of borrowed forms, e.g., *pelati* 'to travel' from Italian and *topor* 'ax' from Turkish. Greenberg points out, "[n]ewer work has shown that *peljati* is native Slavic", as is *topor* (186). In his discussion of morphology, Pavel notes that most village names in Prekmurje are plural feminine nouns ending in *-ce*. Greenberg clarifies that there are still some village names in the west of Prekmurje with feminine plural forms in *-ce*, "[h]owever, the feminine nominative forms in Prekmurje adduced by Pavel have been replaced by masculine plurals in *-ci*" (56).

The translation of Pavel's *Prekmurje Slovene Grammar* is a valuable addition to the available material about a dialect and literary language in an important contact zone at an important time in the development of Slovene. Greenberg's introduction and annotation make the grammar more valuable because they provide broad context and specialized knowledge.

Grant H. Lundberg
Brigham Young University
Provo, UT, USA
grant_lundberg@byu.edu

Tomasz Kamusella. *Politics and the Slavic Languages*. London, New York: Routledge, 2021. xvi + 334 pp. [Series: Routledge Histories of Central and Eastern Europe.] ISBN 978-0-367-56984-6 (hbk); 978-0-367-56985-3 (pbk); 978-1-003-10018-8 (ebk).

Reviewed by Anita Peti-Stantić

The book *Politics and the Slavic Languages* written by Tomasz Kamusella is published in the respected Routledge series, Histories of Central and Eastern Europe. As stated in the preface of the series, “the nations of Central and Eastern Europe experienced a time of momentous change in the period following the Second World War”, but also later, during the Hungarian uprising and the Prague Spring, to name just the most prominent ones, as well as during the tumultuous 1990s and onwards. Therefore, as the editors underline, “the volumes in this series will help shine a light on the experience of this key geopolitical zone and offer many lessons to be learned for the future”.

Kamusella’s monograph consists of seven full chapters preceded by an introduction and followed by an addendum without a commentary (in which the original text of the *Declaration on the Common Language*¹ which circulated in four ex-Yugoslav countries in 2017 is published along with its translation into English) as well as a rather lengthy postscript on methodology. It deals with Slavic languages, especially the politics of the changing number of Slavic languages in the past two centuries, and argues that the politics of language is the politics in Central Europe. The author’s starting point is well known and generally accepted in sociolinguistics—that languages are artifacts and not only systematic entities. Their buildup consists equally of their history and culture, repositories of texts written in specific language varieties in a particular time period and, when it comes to national or official languages, actions of more or less enlightened decision-making bodies. Exactly because of that, and because humans are their creators, they can split or merge the languages according to political or other reasons. Kamusella characterizes these splits and mergers from the point of view of ethnolinguistic nationalism, arguing that the idea of national (official) languages directly corresponds to the splits

¹ Originally presented as *Deklaracija o zajedničkom jeziku* [The declaration on the common language], available at <http://jezicinacionalizmi.com/deklaracija/> (accessed 22 March 2024).

or mergers of nation-states. He even claims, referring to the use of writing systems, spelling conventions, and diacritics for creating and maintaining required differences among the Slavic languages: “These differences are often metaphorically referred to as language frontiers or lines of separation, which keep languages away from one another. These territorializing metaphors are a reflex of the main purpose for which languages are employed in today’s Central Europe—that is, for constructing, legitimizing, and maintaining ethnolinguistic nation-states.”

The monograph comprises the following chapters: 1. A brief *unnatural* history of languages in Europe with several subchapters: What is language?, A brief *unnatural* history of languages in Europe, One Slavic language or three Slavic state languages (and counting)?, Vanishing and metamorphoses, Breakups, Classifying Slavic languages; 2. Nonstate (minority or regional) Slavic languages; 3. The internet: A new frontier; 4. The politics of script; 5. Pluricentric or monocentric; 6. Russian as a pluricentric language; 7. Conclusion: the dilemma of numbers; 8. Addendum: *the Declaration on the Common Language*; 9. Postscript on methodology: People say what they want. As expected from a proper scientific text, the monograph is furnished with an extremely detailed bibliography and index.

Although the proclaimed aim of the book is to cover the past two hundred years of Slavic languages in Europe, most chapters span the recent past or contemporary relations and events. An overview is given in the first chapter, while an explanation of concepts, mostly the ideology of ethnolinguistic nationalism as a way to statehood formation, legitimation, and maintenance in Central Europe, is given in the postscript on methodology. This last chapter ends in a chart 60 pages long of what the author perceives as the formation of what he calls *Einzelssprache/Einzelsprachen* (language/discrete languages). In listing an extraordinary amount of data connected to the selected points in time and space, he starts in the 21st-century postcommunist and cyberspace age and goes back to the Roman Empire and 2nd century BC. In doing so, Kamusella defends his idea that “under the influence of the aforementioned ‘modernizing’ reforms, which were to preserve the existence of non-national polities in Europe and of colonies in Africa and Asia, the estate and colonial structure of society [...] began undergoing transformation. [...] National activists in Europe carried out the program of building nation-states across the continent, while anticolonial-cum-national activists in Africa and Asia adopted this Western ideology as their own and carried out decolonization in line with the logic of this ideology (cf. Mishra 2012). As a result, nationalism became the globe’s first ‘infrastructural ideology’—in other words, the sole accepted ideology of statehood construction, legitimation, and maintenance”.

Since the main motivation for writing this book is the existence of the *Declaration on the Common Language*, I will concentrate on the issues related to concepts central to this impulse. In doing so, I first want to share an

impression backed up by my close reading, to wit, that the book was created, in a way, *backwards*. The author was intrigued by the *Declaration* and seemingly sympathized with its views and its cause, which led him to develop the central ideas in this book. My conclusions come from my reading but also from an analysis of the Index, where a few concepts stand out due to their frequency of appearance. Alongside more general terms, the most prominent concepts used throughout the book are: Einzelsprache/Einzelsprachen, ethnolinguistic nationalism, and nation-states.

This helps one perceive the rather unusual approach to the subject in the topics and the organization of the chapters because the problems the author deals with are not, at first glance, related to each other. What links them together is the idea of ethnolinguistic nationalism, elaborated in ways similar to those sketched in the *Declaration*. The idea of some clash between the Einzelsprachen and common languages runs through the book as a guiding thought. Taking all this into account, it is interesting that the book is, as the author states, offered "to the new generation of scholars from Central Europe, so that they may dare to peer *beyond* the dogma of the nation and the black box of Einzelsprache". I was intrigued when reading this dedication because it is aimed at scholars from Central Europe and not necessarily scholars doing research on Central Europe, which is *per se* interesting. But knowing other works of Kamusella's (especially *The Politics of Language and Nationalism in Modern Central Europe*) as well as having insight into the writing and thinking of most scholars *from* Central Europe, one better understands the need for such a direct call. In my opinion, insisting on ethnolinguistic nationalism as the sole cause of all the ills associated with language formation in the last 200 years, without putting individual linguistic situations and communities in the broader context of European history as well as one's own history and the history of values which persisted in communities for much longer periods (see Katičić 1992), sheds light on only one variable in the political existence of (Slavic) languages in Europe today and therefore asks for critical reading and open discussion.

The writing style is embellished, with many metaphors and with comparisons to the material world aimed at showing how linguistic science differs from the physical sciences, primarily because of the human intervention, which obviously follows the author's thought process. Therefore, I read this book as a combination of a scholarly, argued essay on topics connected through one central idea and as a personal statement on the state of affairs of different Slavic languages. In reading it with that second lens in mind, I agreed with many assertions in the book but also found many points of disagreement, especially where statements were presented without sufficient data and arguments. Therefore, if Kamusella's monograph is to be read as an open text and an invitation to dialogue, I welcome it, albeit with a bit of concern that the success of its reading will depend on the readers'

previous knowledge on the topic, assumptions, and preconceived notions. If, however, the book is to be read mostly by non-specialists in order to inform them of what the author takes to be the truth (as advertised on the publisher's web page) as well as by the new generation of (younger) Slavic scholars to inform them how to correctly treat contemporary Slavic languages, I would be more cautious. That is, although opening some new avenues that have not been systematically discussed in connection with Slavic languages, such as the influence of the internet on the status and corpus of languages, Slavic included, the author puts some questions in the foreground while not even mentioning others.

One of these is the question of how many languages we should even be talking about. On p. 175, he states: "This 'total number of Slavic languages' is a moveable feast, fully depending on the perceiver, alongside changing group views on and attitudes toward what should count as a 'proper' language and what ought not to. During the past two centuries, the pendulum has swung widely from one extreme to another, from a single or just a couple of Slavic languages to many". It is not true that the number of Slavic languages is fully dependent on the perceiver in the way Kamusella suggests by using the metaphor of a "moveable feast". More relevant and important than the number are the criteria according to which one divides the languages. These criteria, be they genetic, typological, or sociolinguistic, are not new in the field. Even sociolinguistic criteria, often disputed, should be discussed and explained, especially as new scholars are an intended audience.

The other one is the perception of the *Declaration on the Common Language*, written five years prior to the publication of this book. While it is true that "some prefer to see it as a single Einzelsprache rather than as the officially recognized and separate four post-Serbo-Croatian languages of Bosnian, Croatian, Montenegrin, and Serbian" (p. 175), sociolinguistics always must ask whether some view is, or historically was, representative of the linguistic community. I am aware that the views from the inside and the outside might differ, primarily in their granularity and emotional load. However, in an attempt at objectivity, one must acknowledge that the majority of linguists, as well as "others" who bothered to take an interest in the *Declaration*, perceived it as a text provoked by the political reason of "unification". The main reason is that the *Declaration* was composed and published 25 years after the South Slavic languages based on the Štokavian dialectal base were formally proclaimed as independent and a few years after Croatian became an official language of the European Union. (Croatia and Slovenia are, amongst ex-Yugoslav states, the only two countries to become members of the European Union to date. Croatian automatically became an official language of the EU like all others.) Therefore, the *Declaration* is really an opinion of some with the right to proclaim and defend their view, which is unquestionable in democracies nowadays. Whether these views stand up to scrutiny through

the lenses of concepts such as ethnolinguistic nationalism (seen as unwanted) and pluricentric languages (seen as wanted) is another question entirely.

For the sake of building a fair argument, I want to remind readers of two sociolinguistic frames. Both can be seen as political, but this is also not new in sociolinguistics (see, for example, Haugen 1983; Joseph and Taylor 1990; Spolsky 2004; Langston and Peti-Stantić 2014). Firstly, by mentioning Anderson's (1983) imagined communities, the author seems to forget that a community, no matter what its size and level of formality, is most often built around some values that shape its identity and that this identity can also be seen as a moving target. Therefore, as there are some who see their language as being a variety of a common language, there are certainly others who see it as a separate one. For some, the main criterion is mutual understanding, while for the others it is the depth of this understanding and the fact that the languages are embedded in culture and literature. Secondly, the concept of pluricentric language is not as undebated as presented in this book. To point only to one issue related with it, one should notice the essential difference between the pluricentric languages listed in the *Declaration* (German, English, Arabic, French, Spanish, Portuguese, and *many other languages across the world* (highlighting APS)) (p. 179 of the reviewed book) and the South Slavic situation. The group of languages that serve as a comparison and point of departure for the formation of a definition are among the most spoken languages in the world, colonial languages spreading over vast territories by means of political power. Therefore, their pluricentricity is of a different kind and guided by different reasons in time and space than, supposedly, the South Slavic ones (see Jakobson 1995 on time and space). For that reason, as well as others, including a viewpoint that there is no significant (or even some) "linguistic segregation and linguistic discrimination in education and public institutions", (p. 180 of the reviewed book) which was one of the main motives for writing the *Declaration*, many members of South Slavic linguistic communities simply considered the *Declaration* as irrelevant, scholarly unargued, and politically incorrect.

To conclude on a positive note and seeking to emphasize the value of the monograph, I want to stress that the author presents many thought-provoking insights and analyses, especially in comparisons with non-Slavic languages. One such instance is when dealing with the nonstate (minority or regional) Slavic languages, where many data points for contacts between domicile and migrant communities are given. Also, he addresses already established issues such as the question of when it is appropriate to treat languages as pluricentric as opposed to monocentric, although the author shifts his attention to Russian from the usual focus on South Slavic languages. This portion of the analysis became even more provocative in the course of events during the past several months, as what Kamusella already classified in the book as "Russia's

continuing ideological and military attack on Ukraine since 2014" (p. 148) reached its apex. I am hopeful that by reading, discussing, and clarifying our positions, our communities can reach a better understanding of and tolerance toward each other in the near future. I see this book as a brick in building this bridge towards helping each other to understand and appreciate each other's values more deeply.

References

- Anderson, Benedict. (1983) *Imagined communities: Reflections on the origin and spread of nationalism*. London: Verso.
- Haugen, Einar. (1983) "The implementation of corpus planning: Theory and practice". Cobarrubias and Fishman, eds. *Progress in language planning: International perspectives*. Berlin, New York, Amsterdam: Mouton, 269–89.
- Jakobson, Roman. (1995) *On language*. Linda Waugh and Monique Monville-Burston, eds. Cambridge (MA): Harvard University Press.
- Joseph, John Earl and Talbot J. Taylor, eds. (1990) *Ideologies of language*. London, New York: Routledge.
- Kamusella, Tomasz. (2009) *The politics of language and nationalism in modern central Europe*. Basingstoke, New York: Palgrave Macmillan.
- Katičić, Radoslav. (1992) *Novi jezikoslovni ogledi*. [New linguistic essays.] Zagreb: Školska knjiga.
- Langston, Keith and Anita Peti-Stantić. (2014) *Language planning and national identity in Croatia*. Basingstoke: Palgrave Macmillan.
- Mishra, Pankaj. (2012) *From the ruins of empire: The revolt against the West and the remaking of Asia*. London: Allen Lane.
- Spolsky, Bernard. (2004) *Language policy*. Cambridge, UK/New York: Cambridge University Press.

Anita Peti-Stantić

Odsjek za južnoslavenske jezike i književnosti

Filozofski fakultet Sveučilišta u Zagrebu

Zagreb, Croatia

anita.peti-stantic@ffzg.hr

Katarzyna Bednarska, Dorota Kruk, Borislav Popov, Olga Saprikina, Traci Speed, Kamil Szafraniec, Svitlana Terekhova, Radislav Tsonev, and Aneta Wysocka, eds. *Contributions to the 23rd Annual Scientific Conference of the Association of Slavists (Polyslav)*. Wiesbaden: Harrassowitz, 2021. 404 pp. [Die Welt der Slaven, 68.] ISBN 978-3-447-11725-8.

Reviewed by Donald F. Reindl

The book *Contributions to the 23rd Annual Scientific Conference of the Association of Slavists (Polyslav)* is a volume of conference proceedings. The Polyslav group was established in 1997 at the University of Konstanz, and it has held annual conferences since then. The group was originally dedicated to sharing research in Slavic linguistics by German-speaking Slavic specialists, and since then it has expanded to encompass a more international scope (Polyslav 2014). The 2019 conference was the last to be held in person for two years; the next two conferences were held online due to the COVID-19 pandemic. In-person Polyslav conferences resumed in 2022.

The volume contains 46 papers presented at the 23rd conference of the Polyslav group from September 9th to 11th, 2019, at Neofit Rilski Southwest University in Blagoevgrad, Bulgaria (Polyslav 2019). Altogether, 72 papers were presented by 75 scholars at the conference, and so the material presented in the volume represents about two-thirds of the contributions from the conference. The articles selected for the publication, which cover a broad variety of topics, underwent a double-blind review process involving 67 reviewers. The articles are generally short, averaging just under nine printed pages each. This reflects their origin as conference papers because the presentation slots were limited to 30 minutes each (Polyslav 2019).

In terms of affiliation, the greatest number of contributors to the volume (22, or just under half) are connected with Polish institutions. This is followed by authors affiliated with institutions in Bulgaria (14, or nearly one-third), Russia (3),¹ Serbia (2), and the Czech Republic, Germany, Japan, Slovenia, and

¹ Contributions by authors affiliated with institutions in Russia are included in this review for statistical purposes only; they are otherwise excluded from commentary due to the ongoing Russian war of aggression against Ukraine. This choice is exclusively based on the contributors' institutional connections and has no implications regarding the personal stances of the authors. This decision is solely the choice of the author of this review, and it is not connected with the opinions or beliefs of the

Ukraine (1 each). The distribution of the languages that the contributions are written in is roughly similar, with Bulgarian (15) and Polish (14) predominating, followed by Russian (8), Ukrainian (3), English and Serbian (2 each), and German and Slovak (1 each). Among the authors using a language that does not match their country of affiliation, the Polish contributors are the most diverse, with articles written in English, Russian, Serbian, and Ukrainian. The articles are arranged in the volume in alphabetical order by surname of the first author rather than in any thematic or other grouping. All of the articles are accompanied by abstracts and keywords in English.

Several of the articles investigate phenomena in more than one language. As one could expect based on the authors' affiliations and the languages of the contributions, articles addressing aspects of Bulgarian (16) and Polish (15) dominate. These are followed by papers that examine or compare Russian (8), Slavic in general and Ukrainian (3 each), Serbian and Slovenian (2 each), and, finally, Balkan languages in general, Belarusian, Czech, English, German, Latin, the Podlachian dialect of Polish, Slovak, and Soviet Romani (1 each).

The diversity of the topics addressed by the articles is vast, but they can be grouped into some common (and occasionally overlapping) categories. The largest thematic group of articles is dedicated to lexis: 13 contributions deal with lexical aspects of language, covering topics such as borrowing, word formation, individual parts of speech (articles, prepositions), onomastics (in particular, nicknames), or vocabulary belonging to specialized semantic areas (ethnonyms and mythology). This is followed by five articles addressing various aspects of language acquisition (including issues related to bilingualism, second or foreign language acquisition, and children's creativity). Four of the papers in the volume are dedicated to morphology, especially word formation (also connected with onomatopoeia), prefixation (including reduplication), and postfixes. Another four of the articles are concerned with syntax (complementation, reduplication and ellipsis in colloquial speech, complex sentences containing motion verbs, and negation). Yet another four of the texts are studies of literature, examining poetry, manuscript tradition reflected in printed works, a contemporary prose writer, and a 17th-century papal brief. Three of the contributions address Slavic culture (18th-century cultural transfer, mythology, and saints).

The volume also contains several topic areas addressed by only one or two texts. Two of the articles investigate semantic issues (relating to adverbs and metaphor in particular), and another pair of articles examine discourse (anti-immigrant children's literature and 17th-century polemical dialogues). Two articles look at translation—focusing on verb forms and modality on the

one hand, and on Biblical onomastics on the other—and yet another two articles examine language fluency in a medical context (Alzheimer's disease and Down syndrome). Finally, phonology and dialect issues (both in Podlachian), a minority language situation (Bulgarian spoken in Moldova), and typology in relation to evidentiality are each addressed by one article.

The nature of a volume of conference proceedings, especially one as varied and extensive as *Contributions to the 23rd Annual Scientific Conference of the Association of Slavists (Polyslav)*, makes it impossible to comment on all of the contributions beyond the general characterizations above. However, a next-best choice is to summarize and comment on a few of the articles in order to provide at least some insight into the “flavor” of the collection. To this end, I have summarized six articles (written in Polish, English, Russian, German, Bulgarian, and Slovak) on various topics as a “sampler” of the research presented.

Katarzyna Bednarska's article “Czym skorupka za młodu nasiąknie. Analiza słoweńskiego dyskursu o migrantach na przykładzie bajki *Deček Anže brani vas Svetje*” (As the twig is bent, so is the tree inclined: Analysis of the Slovenian discourse about migrants exemplified by the children's book *Deček Anže brani vas Svetje*, 31–38) is a topical critical discourse analysis of a Slovenian children's book related to the 2015 European migrant crisis. It opens by sketching the background of the migrant crisis vis-à-vis Slovenia, which was responsible for maintaining a Schengen border, and public reaction to the concept of “securitization” that appeared in public discourse. It was in this context that the magazine *Demokracija* launched a competition in 2018 for an “original Slovenian fairytale”. *Demokracija* is a conservative-to-right-wing publication whose parent company, Nova obzorja, is majority-owned by a Hungarian media company with ties to Prime Minister Viktor Orbán (Košak 2018). The winning story, *Deček Anže brani vas Svetje* (The boy Anže defends the village of Svetje), revolves around a lad that protects his village from “dark men who are evil and wish bad things for Slovenians” (D. F. 2018). Bednarska presents the controversy that this engendered in public debate, contextualizes this new hero with the traditional Slovenian boy-hero Kekec, and draws cogent parallels with notorious examples like Julius Streicher's antisemitic children's book *Der Giftpilz* (The toadstool; 1938). Her analysis covers construction of a threat and the application of propaganda techniques, which she characterizes as an us-versus-them discourse found not only in Slovenia, but throughout Europe and beyond. By deconstructing this example of an anti-immigrant narrative, her article makes a welcome contribution to research on nationalist propaganda in western culture.

Robert Grošelj's contribution “Bulgarian Past Future in Slovene Translations” (117–22) is a contrastive study viewed through the lens of translation. It is of particular interest to Slavic linguistics because it compares two languages that, although they both belong to the South Slavic group, differ

radically in typological terms, especially with regard to the verbal system. The Bulgarian past future (or future in the past), which has no equivalent in Slovenian, is formed by combining the imperfect past of the auxiliary verb *šta* 'will, want' plus the particle *da* 'to' with the present tense of the main verb: for example, *štjaj da vzema* 'I would take' (Lindstedt 1985: 73). It expresses actions that were to be completed in the past but were future from the perspective of another past action, and it has been variously interpreted as an indicative verbal tense or as a Balkan type of conditional (Nicolova 2017: 444), sometimes translated as 'I was on the point of...' (Sussex and Cubberley 2006: 242) or 'I was about to...' (Hauge and Tisheva 2006: 175). After describing the form, Grošelj surveys the modal uses of the construction. He then analyzes how the construction was translated into Slovenian in three novels, which yielded 149 examples. The result is a broad variety of verbal forms in the target language depending on the source-language function of the construction (future in the past, impossibility, possibility, and guessing or wondering): the Slovenian future (sometimes with the desiderative *naj* 'should'), perfect, present conditional (sometimes with *naj*), present, past conditional, and a predicative construction. All in all, the study provides not only a concise inventory of the expressive power of the Bulgarian future past, but also insight into the vast array of choices a translator faces when considering just one verbal form in a related language.

Michał Kozdra's article "Principy leksikograficznego opisanija kulinarnoj leksiki v Učebnom tematičeskom slovaru rusko-pol'skix leksičeskix paralelej" (The principles of the lexicographic description of culinary lexis in *The Learner's Thematic Dictionary of Russian–Polish Lexical Parallels*; 206–15) combines lexicography with synchronic and comparative methods to create an intriguing contrastive presentation of the thematically limited field of culinary arts. The article is based on the first volume of *Dydaktyczny słownik tematyczny rosyjsko-polskich paraleli leksykalnych* (Didactic thematic dictionary of Russian–Polish lexical parallels), which was published in 2019 and is dedicated to culinary material. The dictionary in question is an innovative work that combines lexical parallels between Russian and Polish in a single terminological system: that is, not only false friends, but also words whose meanings match (fully or partially) and international words. The dictionary is aimed at students of the languages and translators, and it draws attention to homonymy and polysemy, as well as stylistic and grammatical differences between such parallel vocabulary (Dubichynskyi 2020). The author describes how the definitions for the dictionary were built, relying on various dictionaries as well as corpora, search engines, various websites, and the author's intuition. The individual entries are divided into thematic groups (dishes, pastries, mushrooms, grain products, dairy products, etc.). The lexemes are then categorized along a gradient of full to partial graphic and phonetic matches (e.g., from *mak/mak* 'poppy' to Rus *salat* 'salad; lettuce'/Pol

salata ‘lettuce’) as well as regular correspondences in terms of suffixation (e.g., *gribok/grzybek* ‘small mushroom’), polnoglasie (e.g., *gorox/groch* ‘pea(s)’), and other characteristics. When the definitions are provided, the result is an invaluable resource for avoiding pitfalls or mistaken assumptions based on parallel vocabulary; for example, *gorčica* refers to both the mustard plant and the condiment, but *gorczyca* is only the plant, and Rus *ukrop* ‘dill’ is etymologically and semantically unrelated to Pol *ukrop* ‘boiling water; heat wave’. The end product is a satisfying work of significant utility, and it can be hoped that other lexicographers will be inspired to create similar dictionaries of lexical parallels.

Tatjana Kurbangulova’s study “Napravo dlja do Solnca: Die Verwendung von Präpositionen in der Herkunftssprache Russisch in Deutschland” (Napravo dlja do Solnca: The usage of prepositions in Russian heritage language in Germany; 216–26) examines heritage speakers’ use of prepositions in Russian. After an overview of the concept of heritage speaker and the function and classification of prepositions, she reviews previous studies on preposition usage by bilingual children and heritage speakers of Russian. Her own study was carried out as part of the project “Russian and Polish Language of Origin as a Resource in School Instruction” and examines material gathered from 11 children age 12 to 14 living in Hamburg. Using oral tests, she collected an extensive corpus (15,073 tokens with a total of 1,097 prepositions). She first uses this material for frequency comparisons with Russian corpora, drawing attention to anomalies, and then she examines the actual usage of the prepositions and their associated cases in greater detail. The participants chose the correct preposition at a rate of 81.9%, with errors such as substitution (e.g., *s Germanija* instead of *iz Germanii* ‘from Germany’) and overuse (e.g., *s mjačikom* instead of *mjačikom* ‘with a ball’), and they used prepositions with the correct cases 76.5% of the time, with the majority of errors involving use of the nominative after the preposition (as in *s Germanija* cited above), as well as frequent confusion between directional and locative functions for prepositions that can take multiple cases. The difficulties that not only foreign learners of Russian experience with prepositions but also some native speakers are notorious, as encapsulated in the title of Terence Wade’s (1982a, 1982b) classic article “Akh, uzhe eti predlogi!” (Oh, those prepositions!). A wide variety of studies have examined not only prepositions, but also other aspects of language use in heritage Russian (cf. Ivanova-Sullivan 2008; Mikhaylova 2012; Polinsky 2008), and this study is a welcome contribution to this growing body of literature.

Kenta Sugai’s article “Săvremennata ezikova situacija v Parkan, Moldova” (The contemporary language situation in Parcani, Moldova; 343–52) investigates the language situation in a Bulgarian-speaking village in southeastern Moldova. Parcani is located in the breakaway region of Transnistria (currently under Russian occupation), and the large majority of its

residents are ethnic Bulgarians, descended from colonists that settled there in 1803 and 1804. Parcani is considered the largest Bulgarian village in Moldova (Grek and Červenkov 2005: 124, 174). This sociolinguistic study is based on fieldwork that the author carried out in the village between 2012 and 2019. Following a presentation of the geographical location of the village, its current ethnic composition (Sugai cites a figure of 81% for Bulgarians, followed by Russians, Ukrainians, and Moldovans), and its settlement history, the author examines the current linguistic situation in Parcani. He shows that, despite the village's overwhelming Bulgarian ethnic majority and its location in Moldova, social conditions have resulted in Russian assuming the role of the dominant language in the official sphere (in particular, for interethnic communication), whereas Bulgarian, Ukrainian, and Moldovan (i.e., Romanian) are relegated to the informal sphere. The article includes supporting images of text in a variety of functions (signs on institutions, public announcements, infrastructure, memorials, and graffiti) to illustrate its findings. The author concludes that the language situation is not only diglossic, but also exoglossic and typologically unbalanced with regard to the unequal status of the high- and low-prestige languages in the community, all of which point to a future tendency to weaken the Bulgarian tradition in the village. The study is a valuable contribution to the literature on the linguistic situation of minority exclaves, which are increasingly vulnerable in the face of globalization.

Jasna Uhláriková's contribution to the volume, "Emocionálne koncepty v slovenskej somatickej frazeológii" (Emotional concepts in Slovak somatic phraseology; 362–70), is an analysis of idioms containing lexemes that refer to parts of the body to express emotions. It takes the model of six basic emotions developed by the psychologist Paul Ekman as a starting point to sort approximately one hundred Slovak idioms collected by the author containing somatic lexemes (e.g., *srdce* 'heart', *koža* 'skin', *noha* 'leg', etc.). She draws on a wide variety of previous literature to present the topic, ranging from linguistic studies such as the seminal work of Lakoff and Johnson to psychological theory. After discussing emotions and phraseology, the article presents typical metaphorical concepts for the six basic emotions (e.g., *radost' je SVETLO* 'joy is LIGHT', etc.) with corresponding Slovak examples. Uhláriková determines that the most frequent body parts constituting such idioms are *oko* 'eye' and *srdce* 'heart', followed by *ústa* 'mouth', *pery* 'lips', and other body parts, and that the most frequent basic emotion expressed by somatic idioms is anger and the least frequent is disgust. Interestingly, some idioms are ambiguous (e.g., *vyskočiť z [vlastnej] kože* 'to jump out of one's skin', which may express joy or anger). The study of metaphor and idioms has a rich tradition in linguistics. Because much of such research concentrates on English, this article—focusing on a less-studied Slavic language—makes an interesting addition to the body of works available to the field. In particular, the information it provides not only has value for the domain of linguistic theory, but will also be welcome

for its obvious applied value to students of Slovak, as well as to lexicographers and translators interested in the contrastive value of the material presented.

Regrettably, it is not possible to summarize all the articles in the volume. Dipping into just a few of the many studies presented provides a taste of its content, and this should certainly whet readers' appetites to explore the work further. The great variety of fields and topics covered by the volume means that there is something of interest in it for every linguist—as well as an opportunity for specialists to expand their horizons by browsing through the volume and reading about research that lies beyond their usual concentration.

In terms of its general layout and mechanical quality, the collection was very well prepared. The contributions have a uniform format and structure, which provides the volume with an overall feel of unity. Its shortcomings are few; there are occasional typos (e.g., *3za* on p. 343) or cited works missing from reference lists (e.g., Lakoff and Johnson 2003 on p. 265), but these do not detract from the overall quality. In a few cases, a different approach would have improved the graphic material (e.g., the pie chart in several similar shades of gray on p. 346 should have been redone with textures or reformatted as a bar chart to aid interpretation). Finally, an index to the volume would have been a welcome addition, although indices are understandably rare in volumes of conference proceedings.

All in all, *Contributions to the 23rd Annual Scientific Conference of the Association of Slavists (Polyslav)* will be welcome reading for any linguist—and especially Slavic specialists—eager to sample the menu of topics it serves up.

References

- D. F. (2018) "Slovenska pravljica, kjer junak brani vas pred 'hudobnimi temnimi možmi'" [A Slovenian fairy tale in which the hero defends the village from "evil dark men"]. 24ur.com (21 December 2018). Available at: <https://www.24ur.com/novice/slovenija/izvirna-slovenska-pravljica-napisana-decek-anze-brani-vas-svetje.html>. Last accessed 30 September 2022.
- Dubichynskiy, Volodymyr [Dubičinskij, Vladimir Vladimirovič]. (2020) "Russko-pol'skij tematičeskij slovar' kulinarnej leksiki/tematičeskij slovar' russko-pol'skix leksičeskix paralelej (kulinarnej leksika)" [Russian–Polish thematic dictionary of culinary vocabulary/thematic dictionary of Russian–Polish lexical parallels (culinary vocabulary)]. Nenad Ivanović et al., eds. *Međunarodna naučna konferencija Leksikologija i leksikografija u svetlu savremenih pristupa. Beograd 28–30 oktobar 2020. Knjiga rezimea* [International scientific conference "Lexicology and Lexicography in Light of Contemporary Approaches". Belgrade 28–30 October 2020. Book of abstracts]. Belgrade: SANU, 9–12.

- Grek, Ivan and Nikolaj Červenkov. (2005) *Po puti nacional'noj duxovnosti Bolgar Moldovy: dokumenty i materialy (konec 80-x – 90-e gg. XX v.)* [Along the path of national spirituality of the Bulgarians of Moldova: Documents and materials (late 1980s–1990s)]. Chişinău: SŞB.
- Hauge, Kjetil Rå and Yovka Tisheva. (2006) *Colloquial Bulgarian*. London: Routledge.
- Ivanova-Sullivan, Tanya. (2008) “Lost in between: The case of Russian heritage speakers”. *Heritage language journal* 6(1): 72–104. Available at: https://brill.com/view/journals/hlj/6/1/article-p72_6.xml?language=en. Last accessed 30 September 2022.
- Košak, Klemen. (2018) “Madžari Janši širijo obzorja” [The Hungarians broaden Janša’s horizons]. *Dnevnik* (29 May 2018). Available at: <https://www.dnevnik.si/1042823887>. Last accessed 30 September 2022.
- Lindstedt, Jouko. (1985) *On the semantics of tense and aspect in Bulgarian*. Helsinki: University of Helsinki, Department of Slavonic Languages.
- Mikhaylova, Anna. (2012) *(In)complete acquisition of aspect in second language and heritage Russian*. Ph.D. dissertation, University of South Carolina.
- Nicolova, Ruselina. (2017) *Bulgarian grammar*. Berlin: Frank & Timme.
- Polinsky, Maria. (2008) “Relative clauses in heritage Russian: Fossilization or divergent grammar?” Andrei Antonenko, John F. Bailyn, and Christina Y. Bethin, eds. *Formal Approaches to Slavic Linguistics #16: The Stony Brook Meeting 2007*. Ann Arbor: Michigan Slavic Publications, 333–58.
- Polyslav. (2014) “About Polyslav: The history of Polyslav”. Available at: <http://www.polyslav-as.org/en/about>. Last accessed 30 September 2022.
- . (2019) “Programme of Polyslav XXIII”. Available at: <http://www.polyslav-as.org/en/conferences>. Last accessed 30 September 2022.
- Sussex, Roland and Paul Cubberley. (2006) *The Slavic languages*. Cambridge: Cambridge University Press.
- Wade, Terence. (1982a) “Akh, uzh eti predlogi! (Problems of prepositional usage in contemporary Russian): 1”. *Journal of Russian studies* 43: 24–32.
- . (1982b) “Akh, uzh eti predlogi! (Problems of prepositional usage in contemporary Russian): 2”. *Journal of Russian studies* 44: 19–28.

Donald F. Reindl
 Department of Translation, Faculty of Arts
 University of Ljubljana
 Ljubljana, Slovenia
donald.reindl@guest.arnes.si

Jan Fellerer. *Urban Multilingualism in East-Central Europe: The Polish Dialect of Late Habsburg Lviv*. Lanham, Maryland: Lexington Books, 2020. viii + 295 pp. [Studies in Slavic, Baltic, and Eastern European Languages and Cultures.] ISBN 978-1-4985-8014-4.

Reviewed by Robert A. Rothstein

Jan Fellerer is associate professor in non-Russian Slavonic languages at Wolfson College of the University of Oxford. His *Urban Multilingualism* is a masterful demonstration of how one can conduct a sociolinguistic study without direct access to speakers of the language(s) in question. His portrayal of what he calls “Lviv borderland Polish” (LBP) is based on close reading of two kinds of material: Polish and Ukrainian popular satirical periodicals and, perhaps surprisingly, police and court records. His reading is informed by his profound knowledge of Polish and Ukrainian, both their standard versions and geographically relevant dialects.

The core of the book consists of four chapters: 1. The City’s Languages, 2. Patterns of Bi- and Multilingualism, 3. Distinct Morpho-Syntactic Characteristics of Lviv Borderland Polish, and a final summary chapter. The languages considered are Ukrainian (especially the Dnister subdialect of its southwestern dialect), Polish (especially LBP, in contrast to the literary standard) and Yiddish (in what he labels as the mid-eastern dialect of eastern Yiddish). German also plays a role as the traditional language of bureaucracy. Unfortunately, although the author mentions *passim* Yiddish influences, he apparently had no sources for examples. (Page 259 offers “Further Yiddishisms are *bajojkis* ‘scrape, unfortunate situation’, Yid. *paihe...*” but this is a non-existent Yiddish word, cited from a source that is absent from the bibliography. That sentence also provides a presumably authentic example from an 1883 issue of the Polish “satirical-political” monthly *Szczutek*: “*zmyszyginowaciac* ‘to go crazy’, Yid. *meschugene*”, which could likely only be from Lviv, where the magazine was published.) The second chapter provides an account of who spoke what in what circumstances. The third chapter is the most detailed,

together with the conclusions constituting nearly 60% of the volume. Finally, there is an impressive bibliography of primary and secondary sources.

Robert A. Rothstein
10 Red Fox Ln
Amherst, MA 01002
USA
rar@slavic.umass.edu

Predrag Piper, Ivan Klajn, and Rajna Dragičević. *Normativna gramatika srpskoga jezika* [Normative grammar of the Serbian language]. 4th revised and enlarged edition. Novi Sad: Matica Srpska, 2020. 766 pp. ISBN 978-86-7946-377-7.

Reviewed by Danko Šipka

Speakers of English may marvel at the word “normative” in the title of this grammar. This word simply rolls off the tongue of Slavic linguists, in sharp contrast to their English-speaking colleagues. Indeed, the single most important difference in ways of maintaining the standard language variety between the English-speaking world and the world of the Slavs is the amount of public prominence given to linguistic norms and those who prescribe them. In the English-speaking world, the standard language variety is maintained by an army of editors, copy editors, language teachers, and others, an army without generals, which tacitly implements the norm. In the world of the Slavs, the spotlight is on the generals—linguists who prescribe linguistic norms and offer normative advice on the product of their work: the norm, as they call it. Serbs are no exception, and this grammar, *Normativna gramatika srpskoga jezika* [Normative grammar of the Serbian language], is proof of the pudding.

The grammar is normative because its authors provide notes to guide users in matters of the linguistic and epilinguistic norms of the standard language variety. For example, when discussing the use of the letter *đ*, the authors note that it is inappropriate to replace it with *dj* (which is a widespread non-standard practice). To provide another example, when discussing the comparative form of adjectives, the authors note that the comparative form of the adjective *visok* ‘tall, high’ is *viši* ‘taller, higher’ rather than *višlji* or *visočiji* (which are common non-standard forms). I have counted well over 500 of such notes throughout the text.

The prominence of this grammar in Serbian culture can be seen in the fact that it has been endorsed by major cultural and political institutions of this ethnic group. The initiative to pen the grammar came from the Serbian Language Standardization Board, a panel of linguists representing major universities, academies of science, and other Serbian cultural institutions (in Serbia and neighboring countries where Serbian is used). The grammar was published by Matica Srpska, the most prestigious Serbian cultural association. Two ministries of the Serbian government funded the work on this grammar

and its publishing. As can be seen, there is formidable firepower behind these linguistic generals, but that is not all.

The first edition of this grammar was authored by Predrag Piper and Ivan Klajn. Both these linguists were academicians of the Serbian Academy of Sciences and Arts and professors at the University of Belgrade (which completes the circle of most prestigious Serbian cultural institutions). This first edition was published in 2013; the second revised and expanded edition followed in 2014. The third edition was the ijekavian version of the second edition (i.e., the version for Serbian speakers in Bosnia and Herzegovina, Montenegro, and Croatia). The present volume is the fourth revised and expanded edition of this grammar, which also includes a third author, Rajna Dragičević, a professor at the University of Belgrade, who mostly contributed to the section about lexical morphology (i.e., word formation). In addition to the expansion of the lexical morphology section, major new reworking in this fourth edition is also to be found in the section on syntax, penned by Predrag Piper. Sadly, Klajn (1937–2021) and Piper (1950–2021) are not among us anymore.

The present grammar is composed of the ingredients that one typically expects in a grammar of a Slavic language. It commences with an introduction (pp. 5–16) and next discusses the script (17–20) and phonology (21–32). Morphonology (morphologically conditioned phonological alternations) is next (33–46), followed by inflectional morphology (47–232) and lexical morphology (233–330). Syntax comes at the end (331–634). The volume is equipped with a standard apparatus (references, symbols and abbreviations, and an index).

What is particularly important about this grammar is that the authors see it (as discussed on pp. 7–10) as a tool in a partnership between what the authors call *codifiers* (linguists who establish the norms of the standard language variety), *educators* (those who educate about those norms, e.g., teachers, parents), and *realizers* (all users of the standard language variety). This is a significant paradigm shift compared to the previous tradition of grammarography, where the users of the standard language variety were expected to obey linguistic authorities without any questioning. While one can question if such a partnership is actually in place, this change in attitude toward democratization of standard-language maintenance should definitely be applauded. The place of this book in a bicentennial history of Serbian grammars and how it relates to the previous tradition is outlined in Šipka 2021.

The presentation of grammatical material in this book follows a traditional structuralist model (e.g., phrase structure and basic clausal analysis in syntax), which means that it continues an established trend in Serbian grammarology (as seen in other recent monolingual grammars, such as Stevanović 1986, 1989; Stanojčić and Popović 1992; and Klajn 2005). This is a logical solution, given that this approach to grammar is taught in schools,

which guarantees wide familiarity with the presentation of the material. As is typical in grammars of this type, the text is segmented in paragraphs, which are often cross-referenced. All grammatical rules are exemplified well, which is also an established tradition in this grammatographical tradition. It is commendable that in their examples, the grammarians use quotes from authors (as a rule, literary writers) far less often than was the case in previous grammars of comparable size (e.g., Stevanović 1986, 1989). Most of the time, this information about the author is sheer ballast for the reader and can be appropriately omitted whenever possible.

This grammar is certain to contribute to fostering Serbian language arts (or, as they are called in Slavic countries, “the culture of language/speech”, e.g., Serbian *jezička kultura*, Polish *kultura języka/językowa/mowy*, Russian *kul'tura reči*). As such, it will go through numerous new editions in the future, which, in turn, offers prospects for its further improvement. In addition to further elaboration and specification of the rules stated in the grammar (which is a matter of course in each new revision of reference works of this kind), the following areas lend themselves to amelioration.

First, the text would benefit from recognizing its target audience and its needs. Right now, there are places where some level of linguistic sophistication is expected from the user and others where everything is explained without any expectations. For example, when discussing vowels, the authors state, “[a]ccording to the horizontal place of formation of sounds in the mouth cavity, the vowels are: front: I, E; central: A; and back: O, U” (23; English transl. D.Š.), without saying that these features rest on the position of the tongue. In contrast, when discussing voiced and voiceless consonants, the authors do explain the physiological background of the feature: “According to the voice parameter, the consonants in the Serbian language are voiced or voiceless. When voiced consonants are pronounced, vocal cords vibrate. When voiceless consonants are pronounced, vocal cords do not participate” (24; English transl. D.Š.). If the audience should include those without linguistic knowledge, the former description needs to be expanded on. If some level of linguistic knowledge is expected from the user, the comment about the vocal cords in the latter description is superfluous.

Second, there is still room for improvement to make the grammar user-friendly. Most of the text is a plain narrative. The text would no doubt benefit from more frequent use of tables and illustrations. For example, a drawing of the vowel diagram would be helpful in the section on phonology, a table summarizing the types of verbal inflection would help in the section on morphology, etc.

Third, the narrative itself could be better streamlined. Quite often, the text reads like an excerpt from an essay, rather than a collection of grammatical rules. Making the description simpler and more impersonal would go a long way toward bringing the text closer to the user. Another aspect of streamlining

would be to separate three content areas that are right now blended in the single narrative: (i) the discussion of general linguistic categories (i.e., what an ending is), (ii) the presentation of the rules of Serbian grammar, and (iii) the commentary about historical and current developments around the rule in question. It is commendable that normative notes have been set out in a separate section following paragraphs to which they pertain. A similar segmentation could be done with the three aforementioned content areas, which are currently intertwined. That segmentation, too, would make the text much easier to follow.

References

- Klajn, Ivan. (2005) *Gramatika srpskog jezika* [Grammar of Serbian]. Belgrade: Zavod za udžbenike i nastavna sredstva.
- Stanojčić, Živojin and Ljubomir Popović. (1992) *Gramatika srpskoga jezika: Udžbenik za I, II, III i IV razred srednje škole* [Grammar of Serbian: Textbook for I, II, III, and IV classes of high school]. 2nd updated ed. Belgrade/Novi Sad: Zavod za udžbenike i nastavna sredstva/Zavod za izdavanje udžbenika.
- Stevanović, Mihailo. (1986) *Savremeni srpskohrvatski jezik* [Modern Serbo-Croatian language]. Vol. 1. 5th ed. Belgrade: Naučna knjiga.
- . (1989) *Savremeni srpskohrvatski jezik* [Modern Serbo-Croatian language]. Vol. 2. 4th ed. Belgrade: Naučna knjiga.
- Šipka, Danko. (2021) “Serbskie (normativnye) grammatiki na sociokul’turnom fone svojego vremeni” [Serbian (normative) grammars against the socio-cultural background of their time]. Motoki Nomati and Siori Kiësava, eds. *Grammatika v obščestve, obščestvo v grammatike: Issledovanija po normativnoj grammatike slavjanskix jazykov* [Grammar in society, society in grammar: Studies on normative grammar of Slavic languages]. Moscow: Izdatel’skij Dom JaSK, 193–218.

Danko Šipka
 School of International Letters and Cultures
 Arizona State University
 Tempe, AZ, USA
 danko.sipka@asu.edu

T. I. Vendina. *Praslovjanskoe slovo vo vremeni i prostranstve Slavii* [Proto-Slavic words in time and space of the Slavia]. Moscow, St. Petersburg: Nestor-Istorija, 2022. 344 pp. ISBN 978-5-4469-2067-9.

Reviewed by Danko Šipka

The book under review represents another important study which draws upon the ultimate treasure trove of Slavic linguistics, the *Slavic Linguistic Atlas*, better known by its Russian abbreviation OLA (*Obščeslavjanskij lingvističeskij atlas*, <https://www.slavatlas.org>). The present study is a welcome contribution to Slavic lexicology and historical, areal, and typological linguistics. Its author, Professor Tatjana Ivanovna Vendina, is a leading global authority on this subject. This particular monograph by her is the latest in a series of important books, starting with *The Differentiation of Slavic Languages Based on Word Formation Data* from 1990, through *Medieval Man in the Mirror of the Old Church Slavonic Language* (2002), to, most importantly and connectedly, *A Typology of Lexical Areas of Slavia* (2014), to mention just the most interesting ones (see more at <https://inslav.ru/people/vendina-tatyana-ivanovna>). Being based on the OLA, this monograph also dovetails with various other recent publications about the lexicon of Slavic languages (Kurkina 2021; Marković 2020; OLA 1988, 2000a, 2000b, 2003, 2007, 2009, 2012, 2015, 2020; OLABG 2015; Saenko 2022, to name just the most immediately relevant ones).

The present monograph is intriguing even in its architecture. In addition to the Introduction and Conclusion, it includes four chapters which zoom in on the words inherited from Proto-Slavic, first by all three groups of Slavic languages, then by two groups, then by one group, and finally by one Slavic language. The book is also equipped with an index of lexemes featured on the maps, the maps (as many as 195 of them), an index of words, a list of references, and a list of places encompassed by the OLA.

The issues that the present monograph addresses put it in dialog with what is arguably the central question of historical linguistics, dialectology, and areal linguistics, namely: what is the relationship between maintenance and shift, as temporal categories, on one hand, and the area encompassed by stability or shift, as a spatial category, on the other hand. In contributing her evidence to elucidation of the broader question of how lexemes exist in the time-space continuum, Vendina deploys sound methodology. She has selected non-derived Proto-Slavic lexemes with Indo-European origin and ex-

plored their preservation and loss in Slavic dialects. The selection criteria as well as methodologically relevant topographic and linguistic criteria are discussed in the Introduction. The author also engages in dialog with previous researchers in this field, such as Bernštejn, Trubačev, and Tolstoj, to name just the most famous ones.

As previously noted, the main narrative part of the book comprises three chapters where the author traces the preservation of Proto-Slavic non-derived lexemes with Indo-European background first in the three groups of Slavic languages (chapter 1), then in pairs of branches of Slavic languages (chapter 2), in each individual branch (chapter 3), and finally in individual Slavic languages (chapter 4). Each of these three chapters details various configurations of preservation, providing ample data about each lexeme (its reconstructed form, meaning, sound shifts, sources of attestation, etc.). Each configuration is also depicted on one of the aforementioned 195 maps in the attachments.

Thus, in chapter 1, we can find various configurations of lexical preservation, from the situation that a lexeme has Slavic-wide distribution, such as **ledъ* 'ice', to those that have a wide distribution in East and South Slavic dialects and a limited distribution in West Slavic, such as **lic-e* 'face', and to those that have a limited distribution in all three groups of Slavic dialects, e.g., **jam-a* 'grave, burial site'. Following this list of configurations, the author lists which of these lexemes appear in the dialects of each of the major Slavic ethnic groups (Russian dialects, Ukrainian dialects, etc.). This serves as a sort of an index showing the number of the map in which each lexeme is treated. Next, the author provides a list of maximally distributed lexemes. The chapter concludes with a discussion of the distribution of various lexical fields (such as the animal world, agriculture, etc.) in Slavic dialects and the distribution of all observed lexemes across Slavic dialects.

Similarly, chapter 2 first provides binary configurations of preservation, from those covering the entire Western and Eastern Slavic dialectal space, such as **slov-o* 'word', to those broadly distributed in West Slavic dialects but with a limited distribution in their East Slavic counterparts, e.g., **vin-o* 'grapes', and those having a limited distribution in Eastern and Southern Slavic dialects, such as **kor-a* 'eggshell', to name just a few configurations. In this chapter, too, the distribution of lexemes in the dialects of major Slavic nations is next, followed by a list of maximally distributed lexemes.

Chapter 3 follows the same script. One can find configurations such as the distribution of a lexeme throughout the South Slavic dialectal space, e.g., **drъv-o* 'tree', a wide distribution in the Czech dialects with a limited distribution in other West Slavic dialects, such as **gonъ* 'hunt', and many others. Here too, the distribution of these lexemes in the dialects of major Slavic ethnic groups and a list of lexemes with maximal distribution follow. Finally, chapter 4 lists configurations by individual languages, such as the words found only in Ukrainian, e.g., **žag-a* 'thirst'.

The Conclusion presents major findings from the four main chapters, which can be summarized as follows. In the South Slavic realm, the highest level of preservation is to be found in most Slovene dialects, in Croatian Kajkavian dialects along the Slovene border, and in Serbian Zeta-Sjenica dialects. In West Slavic, the highest level of preservation is attested in Western Czech, Polish Małopolska dialects, and various Slovak dialects. In the East Slavic group, Ukrainian dialects show the highest level of preservation, most notably the Southwestern and Polesian dialects. The author also contextualizes her findings and points to possible further research in this field.

With this new monograph that continues a brilliant series penned by Professor Vendina and other scholars in and around OLA, we have received a wellspring of lexical information for various studies in the fields of lexicology, dialectology, areal, and typological linguistics. For example, it is worth exploring how well correlated is the level of lexical preservation with the degree of grammatical conservatism. One cannot help noticing that, for example in the South Slavic realm, those dialectal groups that are grammatically conservative also feature an equally conservative lexical stock. Needless to say, one should go a step further and ask what kind of geographical and historical circumstances contribute to a higher level of lexical preservation.

Concluding, one can add that the use of the magnificent wealth of information displayed in this monograph by Vendina and all other publications listed in the references below would be greatly facilitated if they were available as online searchable databases and GIS maps. Sadly, the current atmosphere in Slavic studies, where politics interferes with the profession, does not leave much hope that this brilliant data would reach its fully searchable potential any time soon.

References

- Kurkina, L. V. (2021) *Slavjanskoe slovo vo vremeni i prostranstve*. [Slavic word in time and space.] Moscow: Indrik.
- Marković, Marjan, ed. (2020) *Lingvistički atlas na makedonskite dijalekti*. [Linguistic atlas of Macedonian dialects.] Skopje: MANU.
- OLA. (1988) *Obščeslavjanskij lingvističeskij atlas. Serija leksiko-slovoobrazovatel'naja. Vypusk 1. Životnyj mir*. [Slavic linguistic atlas. Lexical and word-formation series. Volume 1. Animal world.] Moscow: Nauka.
- . (2000a) *Obščeslavjanskij lingvističeskij atlas. Serija leksiko-slovoobrazovatel'naja. Vypusk 2. Životnovodstvo*. [Slavic linguistic atlas. Lexical and word-formation series. Volume 2. Livestock farming.] Warsaw: Instytut Języka Polskiego PAN.

- OLA. (2000b) *Obščeslavjanskij lingvističeskij atlas. Serija leksiko-slovoobrazovatel'naja. Vypusk 3. Rastitel'nyj mir.* [Slavic linguistic atlas. Lexical and word-formation series. Volume 3. Plants.] Minsk: Nacyjanal'naja akademija nauk Belarusi.
- . (2003) *Obščeslavjanskij lingvističeskij atlas. Serija leksiko-slovoobrazovatel'naja. Vypusk 8. Professii i obščestvennaja žizn'.* [Slavic linguistic atlas. Lexical and word-formation series. Volume 8. Professions and social life.] Warsaw: Instytut Języka Polskiego PAN.
- . (2007) *Obščeslavjanskij lingvističeskij atlas. Serija leksiko-slovoobrazovatel'naja. Vypusk 6. Domašnee xozjajstvo i prigotovlenie pišč.* [Slavic linguistic atlas. Lexical and word-formation series. Volume 6. Housekeeping and cooking.] Moscow: Rossijskaja akademija nauk. Institut russkogo jazyka im. V. V. Vinogradova. Institut slavianovedenija.
- . (2009) *Obščeslavjanskij lingvističeskij atlas. Serija leksiko-slovoobrazovatel'naja. Vypusk 9. Čelovek.* [Slavic linguistic atlas. Lexical and word-formation series. Volume 9. Person.] Kraków: Instytut Języka Polskiego PAN.
- . (2012) *Obščeslavjanskij lingvističeskij atlas. Serija leksiko-slovoobrazovatel'naja. Vypusk 4. Sel'skoe xozjajstvo.* [Slavic linguistic atlas. Lexical and word-formation series. Volume 4. Farming.] Bratislava: Jazykovedný ústav Ľudovíta Štúra Slovenskej akademie vied. VEDA.
- . (2015) *Obščeslavjanskij lingvističeskij atlas. Serija leksiko-slovoobrazovatel'naja. Vypusk 10. Narodnye obyčaji.* [Slavic linguistic atlas. Lexical and word-formation series. Volume 10. Folk customs.] Moscow: Nestor-Istorija.
- . (2020) *Obščeslavjanskij lingvističeskij atlas. Serija leksiko-slovoobrazovatel'naja. Vypusk 12. Ličnye čerty čeloveka.* [Slavic linguistic atlas. Lexical and word-formation series. Volume 12. Personal traits of people.] St. Petersburg: Nestor-Istorija.
- OLABG. (2015) *Obštoslavjanski lingvističen atlas. Bălgarski materiali.* [Slavic linguistic atlas. Bulgarian materials.] Sofia: BAN.
- Saenko, M. N. (2022) *Očerki po slavjanskoj somatičeskoj leksike.* [Essays on Slavic somatic vocabulary.] Moscow: Indrik.

Danko Šipka
 School of International Letters and Cultures
 Arizona State University
 Tempe, AZ, USA
 Danko.Sipka@asu.edu

Steven Franks. *Microvariation in the South Slavic Noun Phrase*. Bloomington, Indiana: Slavica Publishers, 2020. 518 pp. ISBN 978-0-89357-498-7.

Reviewed by Aida Talić

The structure of the nominal domain poses many interesting questions for linguistic theory, both from perspectives that focus on individual languages and from cross-linguistic perspectives. Despite the volume of available research on this topic, how much functional structure (if any) is projected in the extended domain of N of a given language and whether all languages have uniform extended domains of N is still largely debatable. Within generative syntax (assumed in this book), before Chomsky's "Remarks on Nominalization" (Chomsky 1970), the structure of the nominal domain was quite different from the clausal domain, the topmost projection of a nominal domain being a lexical projection, NP, with D introduced as its modifier. The recognition that there are some parallelisms between the nominal and clausal domains, as well as the rise of X-bar syntax (Chomsky 1970; Jackendoff 1977), turned the tables, and D became a head projecting a DP layer above NP (Szabolcsi 1983; Fukui 1986; Abney 1987). While, for languages like English, the DP hypothesis is assumed by most linguists, there are also a few who argue against the DP hypothesis altogether (e.g., Payne 1993; Bruening 2009). From a crosslinguistic perspective, many have suggested or adopted the Universal DP Hypothesis, where it is argued that every nominal domain universally projects a DP (Bowers 1991; Longobardi 1994; for Slavic, Progovac 1998; Leko 1999; Pereltsvaig 2007, a.o.). However, certain typological differences between languages that have articles and those that lack articles have led to a parametric approach to the nominal structure, where only languages that have articles project a DP (Bošković 2005, 2008, et seq.; Despić 2011, a.o.). Some later cross-linguistic work further shows that some languages exhibit mixed behavior and are not easily classified within the two-way cut between NP and DP languages, arguing for a three-way typology, where languages with affixal articles represent a middle case between NP and DP languages (Talić 2015, 2017), or for an even more fine-grained scale, where even Italian can either project a DP or have the D head adjoin to N without projecting (Oda 2022). Other than DP, functional structure in the nominal domain in Slavic has also been proposed for other purposes (e.g., Aljović's 2002 analysis of BCMS long-form adjectives involves FPs above NP, where F hosts the long-form inflection). This book contributes essential empirical detail as well

as very appealing theoretical proposals towards an answer to these important questions about the kinds of functional projections present in the nominal domain and which Slavic languages have them. The author zooms in on the microvariation between closely related languages and arrives at a nuanced proposal for the structure of the noun phrases in South Slavic languages. The major levels of projections above NP explored are DP and KP, where South Slavic languages either project all the way up to KP (e.g., Bulgarian), or lack a DP but have a KP (e.g., BCMS), or they are in a transitional stage of language change: Slovenian turning from a KP-only to a KP-and-DP language, and Macedonian turning from a KP-and-DP language to a DP-only language. These proposals are motivated throughout the book by a closer look into microvariation regarding a range of phenomena—clitics, reflexives (e.g., the Bulgarian *nego si* construction), clitic doubling, orphan accusative, agreement in coordinations and agreement with “hybrid” nominals.

The author starts with the background necessary to follow work on microvariation and general syntactic architecture. He summarizes his chosen view of language change as “a failure in the transmission across time of linguistic features” (Kroch 2001) in situations where linguistic input during language acquisition could be analyzed in more than one way, given that this book deals with closely related languages and a seemingly identical sentence in several Slavic languages may have slightly or significantly different structures. A possibility of having two competing structures available in certain situations is also entertained in the book. The standard minimalist bottom-up structure building in line with the Bare Phrase Structure notation (i.e., not indicating X' levels unless there is a clear specifier) is assumed, but the author uses XP labels for clarity, to indicate that a certain head does not project further. For word-building, the Distributed Morphology (DM) framework is adopted, although the use of the term “Vocabulary Item” seems to depart from how it is typically used in the DM literature. That is, the author states that “vocabulary items are constructed not only in the course of the syntax, but also on the PF-side of the grammar” (p. 10), which seems to indicate that the term “vocabulary item” here means something closer to “word”, or a “complex syntactic head”, or an entry in a dictionary of a language, rather than a phonological exponent of an abstract morpheme that is not present in the course of the syntactic derivation and only gets inserted at the PF side (Harley and Noyer 1999). Regarding the presence of functional categories cross-linguistically, the author rejects the universalist approach mentioned above and adopts the view that languages may differ in the amount of functional structure projected above NP (and in other domains) and that meaning alone is not sufficient to motivate structure. Rather, additional morphological or syntactic motivation is necessary to give rise to a syntactic functional projection.

More specific assumptions about features and potential functional projections in the nominal domain in South Slavic are then summarized and

explored. The author discusses two views on features—privative (Harley and Ritter 2002) vs. polar (Halle 1997). The former, adopted in Franks (2017) and in this book (e.g., discussion of agreement in chapter 7), is the view that features are organized into hierarchies and that the presence or absence of individual features leads to specific values (e.g., 1st person = [_{PERS} PART, AUTH], 2nd person = [_{PERS} PART], 3rd person = [_{PERS} \emptyset]), rather than that features have an off-and-on switch but are always present in the feature bundle (as in the latter view). A variety of options are considered for how nominal features like person, number, gender, case, definiteness, etc., are introduced in the derivation and whether they project syntactic functional layers. Regarding the category of pronouns, which are typically treated as Ds, the author addresses a well-known contrast between BCMS and Italian (Progovac 1998), where in the former only pronouns can move higher than the adjective *sama* ‘alone.F’, but in the latter, both pronouns and proper names move higher than *sola* ‘alone.F’ (pp. 48–49). He takes BCMS pronouns to be realized in the head K unlike in Italian (where pronouns and proper names move to D), given that the motivation for this movement in BCMS cannot be definiteness (crucially pointing out that proper names get their definiteness/specificity without moving). More broadly in Slavic, the author argues that the KP projection is present and hosts case features which are valued by a corresponding functional projection in the clausal structure. Regarding definiteness, the author assumes that in languages like English or Bulgarian, this feature projects a DP, while in others it is an auxiliary feature appearing on some other projection (e.g., on KP in BCMS) without resulting in a functional layer of its own (see also Oda’s (2022) treatment of D in Italian, where D can undergo head-adjunction to either N or some higher functional projection). This approach may shed light on some important questions about how definite and indefinite interpretation is achieved in the grammar if languages vary in how much structure they project in the nominal domain. More specifically, a point of frequent criticism of non-universalist approaches to the presence of DP cross-linguistically has been that definiteness is achieved differently in languages with and languages without articles. In the former, the head D introduces the *iota* operator in the course of the syntactic derivation and semantics merely “reads” the interpretation from the composition of D and N, where the *iota* operator turns the property (predicate) into a unique individual having that property (argument) (Partee 1986). In the latter, the *iota* operator is not introduced during the syntactic derivation, so various type-shifting operations are used by semantics to switch between predicates and arguments (Partee 1986; Chierchia 1998). Thus, if the definiteness feature can be present in the syntax even if it does not project its own phrase (i.e., if the *iota* operator can be introduced as a part of a complex syntactic head that also introduces other features), we capture the effects of the lack of DP projection in the syntax, but the semantics still has the same combination of pieces contributed during the syntactic derivation to interpret.

After all, Chierchia's (1998) Blocking Principle ("Don't do covertly what you can do overtly") only applies straightforwardly under lexicalist assumptions where it is known before spell-out whether a language has an overt or covert article, which determines whether or not type-shifting as a last resort can take place. Assuming "late insertion" in line with DM alongside the Y-model of grammar, the availability of an overt exponent for a definite article would not be visible to the LF component, so the Blocking Principle could not be a condition applied at the syntax-semantics interface and would at most be a more general economy condition. Assuming that the definiteness feature is always introduced in the syntax, whether or not it projects a phrase, seems to be more compatible with the late insertion assumption adopted in this book. For gender and number features, several options are considered, both where these features project their own phrases and where they are introduced alongside other features in heads that are not specifically designated to them (e.g., little *n* for Gen and *D* for Num), but the author does not clearly choose one option over the other. However, these projections do not show up in later structures in the book, so the latter option seems to be adopted at least for the sake of simplicity.

The author then addresses some issues in binding posed by the colloquial form *nego si* 'him self' in Bulgarian, a pattern observed by Schürcks (2003, et seq). This exploration leads to proposing a more detailed structure for Bulgarian KP, with an AgrP between KP and DP. Interestingly, this reflexive form is not available in Macedonian, despite the two languages often being classified as having the same nominal structure, as the only two Slavic languages with overt definite articles. This leads the author to propose in chapters 5 and 6 that Macedonian has simpler nominal structure than Bulgarian. The other two languages closely contrasted are BCMS and Slovenian. While there is no separate chapter focusing specifically on nominal functional projections in BCMS, it is hinted throughout the book that this language has a KP to host clitics and help derive full pronominal forms. I wonder if this projection can also be hosting some elements usually classified as prepositions, especially in situations where Genitive case alternates with *od* 'of' in BCMS (see PP-complement extraction cases in Talić 2019: 1133-34). For Slovenian, it is argued based on Orphan Accusatives that this language is in an early stage of developing a DP (between NP and KP), and the author labels this projection IndefP, arguing that this projection is responsible for particular interpretations Orphan Accusatives get, since language change towards DP emerging often starts with the indefinite article. The final chapter addresses agreement in coordinations and agreement with hybrid nominals. While the author reviews key patterns and cross-dialectal variation in these contexts and provides interesting accounts, this chapter seems the least connected to the rest of the book and could have perhaps been left out for a separate project. From the discussion provided in chapter 7, it is not clear how the functional

structures specifically proposed for South Slavic languages discussed in the rest of the book bear on these agreement phenomena.

Overall, for its attention to empirical detail and microvariation both between related languages and dialects, interesting theoretical proposals, and open questions raised throughout, this book is likely to be a stimulating read and resource for researchers investigating the structure of the nominal domain across Slavic and beyond as well as for students searching for topics and open questions in this area. While it is at times difficult to follow what particular set of assumptions are finally adopted for a particular structure, it is commendable how many different options for various portions of the structures are considered throughout the book. This is one of the rare sources that take microvariation and cross-dialectal differences seriously and engage with it, rather than focusing on judgments from the majority of speakers and treating the rest as exceptions to put aside or as noise. Given that language change is an unstoppable force, there is certainly a lot of microvariation that can shed light on many important questions, as was done in this book for the nominal domain.

References

- Abney, Steven Paul. (1987) "The English noun phrase in its sentential aspect". Unpublished Ph.D. dissertation, Massachusetts Institute of Technology.
- Aljović, Nadira. (2002) "Long adjectival inflection and specificity in Serbo-Croatian". *Recherches linguistiques de Vincennes* (31): 27–42.
- Bošković, Željko. (2005) "On the locality of left branch extraction and the structure of NP". *Studia linguistica* 59(1): 1–45.
- . (2008) "What will you have, DP or NP?" *Proceedings of NELS* 36(1): 101.
- Bowers, John. (1991) "The syntax and semantics of nominals". S. Moore and A. Z. Wyner, eds. *Cornell working papers in linguistics* 10: 1–30, Ithaca, NY: Cornell University.
- Bruening, Benjamin. (2009) "Selectional asymmetries between CP and DP suggest that the DP hypothesis is wrong". *Proceedings of the 32nd annual Penn linguistics colloquium (PLC 32)*(1): 28–35.
- Chierchia, Gennaro. (1998) "Reference to kinds across language". *Natural language semantics* 6(4): 339–405.
- Chomsky, Noam. (1970) "Remarks on nominalization". R. Jacobs and P. Rosenbaum, eds. *Readings in English transformational grammar*. Waltham: Ginn, 184–221.
- Despić, Miloje. (2011) "Syntax in the absence of determiner phrase". Unpublished Ph.D. dissertation, University of Connecticut, Storrs, CT.

- Franks, Steven. (2017) *Syntax and spell-out in Slavic*. Bloomington, Indiana: Slavica Publishers.
- Fukui, Naoki. (1986) "A theory of category projection and its application". Unpublished Ph.D. dissertation, Massachusetts Institute of Technology.
- Halle, Morris. (1997) "Distributed morphology: Impoverishment and fission". Benjamin Bruening, Yoonjung Kang, and Martha McGinnis, eds. *MIT working papers in linguistics* 30, Papers at the Interface, Cambridge, MA, 425–49.
- Harley, Heidi and Rolf Noyer. (1999) "Distributed morphology". *Glott international* 4(4): 3–9.
- Harley, Heidi and Elizabeth Ritter. (2002) "Person and number in pronouns: A feature-geometric analysis". *Language* 78(3): 482–526.
- Jackendoff, Ray. (1977) *X syntax: A study of phrase structure*. Cambridge, MA: MIT Press.
- Kroch, Anthony. (2001) "Syntactic change". Mark Baltin and Chris Collins, eds. *The handbook of contemporary syntactic theory*. Malden, MA: Blackwell, 629–739.
- Leko, Nedžad. (1999) "Functional categories and the structure of the DP in Bosnian". M. Dimitrova-Vulchanova and L. Hellan, eds. *Topics in South Slavic syntax and semantics*. Amsterdam: John Benjamins, 229–52.
- Longobardi, Giuseppe. (1994) "Reference and proper names: A theory of N-movement in Syntax and Logical Form". *Linguistic Inquiry* 25(4): 609–65.
- Oda, Hiromune. (2022) "The NP/DP-language distinction as a scale and parameters in Minimalism". Unpublished Ph.D. dissertation, University of Connecticut, Storrs, CT.
- Partee, Barbara. (1986) "Noun phrase interpretation and type-shifting principles". Jeroen Groenendijk, Dick de Jongh, and Martin Stokhof, eds. *Studies in discourse representation theory and the theory of generalized quantifiers*. Berlin, Boston: De Gruyter, 115–44.
- Payne, John. (1993) "The headedness of noun phrases: Slaying the nominal hydra". G. G. Corbett, N. M. Fraser, and S. McGlashan, eds. *Heads in grammatical theory*. Cambridge, UK: Cambridge University Press, 114–39.
- Pereltsvaig, Asya. (2007) "The universality of DP: A view from Russian". *Studia linguistica* 61(1): 59–94.
- Progovac, Ljiljana. (1998) "Determiner phrase in a language without determiners". *Journal of linguistics* 34: 165–79.
- Schürcks, Lilia. (2003) "Binding and Bulgarian". Ph.D. dissertation, Rijksuniversiteit Groningen.
- Szabolcsi, Anna. (1983) "The possessor that ran away from home". *The linguistic review* 3:89–102.
- Talić, Aida. (2015) "Adverb extraction, specificity, and structural parallelism". *Canadian Journal of Linguistics/Revue canadienne de linguistique* 60(3): 417–54.

- Talić, Aida. (2017) "From A to N and back: Functional and bare projections in the domain of N and A". Unpublished Ph.D. dissertation, University of Connecticut, Storrs, CT.
- . (2019) "Upward P-cliticization, accent shift, and extraction out of PP". *Natural language & linguistic theory* 37: 1103–43.

Aida Talić
Department of Linguistics
University of Illinois Urbana-Champaign
Urbana, IL, USA
talicaida@gmail.com

Zrinka Kolaković, Edyta Jurkiewicz-Rohrbacher, Björn Hansen, Dušica Filipović Đurđević, and Nataša Fritz. *Clitics in the Wild: Empirical Studies on the Microvariation of the Pronominal, Reflexive and Verbal Clitics in Bosnian, Croatian, and Serbian*. Berlin: Language Science Press, 2022. 461 pp. [Open Slavic linguistics, 7.] ISBN 978-3-98554-032-7 (hardcover).

Reviewed by Anton Zimmerling¹

The reviewed book offers an empirically oriented description of Bosnian, Croatian, and Serbian (BCS) clitics, with a focus on those features that are subject to parametric microvariation in regional varieties of BCS and across them (p. 5). Descriptive grammars of BCS include a brief mention of BCS proclitics, including the conjunctions *i*, *a*, and the negator *ne* (Browne and Alt 2004: 15), but the authors of the reviewed book restrict their analysis to BCS clustering enclitics representing two kinds of sentence categories—oblique pronouns and auxiliaries. BCS is traditionally described as a language with 2P clitics, where the clustering clitics do not take the clausal left edge ($\# \dots \text{CL} \dots \ast \# \text{CL} \dots$) and behave as strict enclitics, i.e., they always need a *non-clitic* host to their left ($X/XP = \text{CL}$).² The authors confirm this view and state that clitic-first (1P) orders do not occur in any standard regional variety of BCS, though some Neo-Štokavian dialects license clustering clitics after initial proclitics (cf. *I = su.AUX.3PL = ga.3SG.M strelali.PTCP.3PL* ‘and they shot him’), while the Banatsko-pomoriški subdialect, Kosovsko-resavski, Prizrensko-južnomoravski, and

¹ This paper has been written with support from the project “Parametric Description of Languages of the Russian Federation”, realized at Pushkin State Russian Language Institute. I am indebted to Wayles Browne and Jasmina Milićević for their valuable comments. The sole responsibility is mine.

² The strict enclisis phenomena in Romance and Slavic languages are often explained by the so-called Tobler-Mussafia law, i.e., a presumably non-syntactic condition excluding the clustering object pronouns and auxiliaries from the clausal left edge and leaving them in 2P in some clausal types, e.g., in imperative clauses in French, Italian, Romanian, or Macedonian, etc., or in all types of clauses, e.g., in Bulgarian and BCS (cf. Franks 2008, 2017: 188). However, strict enclitics do not necessarily take 2P.

Timočko-lužnički dialects³ license clustering clitics in the absolute initial position (pp. 160–62).⁴

In §2.3 the authors specify that they are interested in “systemic microvariation, which is defined as purely language-internal”, and not in sociolinguistic triggers of variation (p. 14). The reviewer has tried to apply a similar approach to other Slavic languages⁵, although there is no obvious way to eliminate the external factors completely, since all kinds of supra-individual variation are in a broad sense sociolinguistic. This difficulty is confirmed by the high frequencies of some sociolinguistic terms in the reviewed book. The authors adopt the distinction of *diatopic* variation, depending on space; *diaphasic* variation, depending on the modes of language (oral vs. written, standard vs. sub-standard); and *diastratic* variation, reflecting the use of different social groups from the classic work by Coseriu (1980). They use the first two terms consistently. The term “diatopic” occurs 47 times, while the term “diaphasic” is used 57 times. The term “diastratic” occurs only 7 times, which is in accord with the explicit wish of the authors not to deal with language strata (p. 6). However, they state that the inflected forms of the conditional clitics *bih*, *bismo*, etc., in place of the uninflected *bi*, are better preserved in the corpus of spoken Bosnian by more educated speakers, which is then an instance of diastratic variation (pp. 191–92).

The book’s conception is introduced in Chapter 2, entitled “Terms and Concepts in the Light of Theoretical Approaches to the Study of Clitics in BCS”. The authors briefly characterize the notions of “clitic” and “clitic cluster” and state that BCS is a language with “clitic clusters” (pp. 18–22). There are two main aspects of clitic syntax in languages of this class—clitic-internal ordering, i.e., the structure of clitic clusters projected by template rules (Franks and King 2000; Zaliznjak 1993: 282), and clitic-external ordering, i.e., the placement of clusters and clustering clitics in a single clause or complex of clauses. The authors focus on clitic-external ordering, since there seems to be only a limited, dialect-bound variation in the internal organization of

³ The non-translated names of BCS dialects are given here as they are used in the reviewed book (see Figure 7.1 on p. 129 of the reviewed book for a dialectal map and Table 7.1 on p. 130 for alternative dialectal divisions).

⁴ These options displayed by different BCS dialects have different historical explanations. The 1P orders in the Kosovsko-resavski, Prizrensko-južnomoravski, and Timočko-lužnički dialects are likely due to contacts with word-order systems having *v*P-internal clitics, where the Tobler-Mussafia law does not hold, while the proclitic-enclitic complexes in Neo-Štokavian might be a remnant of Proto-Slavic syntax, since some Old Slavic idioms have this feature as well.

⁵ Cf. Zimmerling 2018 and Zimmerling 2022 for Modern Russian and Early Old Russian, respectively, and Ivanova and Zimmerling 2019 for a contrastive analysis of Russian and Bulgarian.

BCS clitic clusters,⁶ cf. the short sections 6.4 and 7.5. For the same reason, the authors do not discuss the syntax of the polar ‘yes-no’-question particle *li*, despite its being part of the BCS cluster—“there is no variation of the CL particle *li* in BCS varieties” (p. 99).

Clitic studies is a research field that calls for a great variety of terms referring to different dimensions of clitic classification; cf. “simple clitics” vs. “special clitics” (Zwicky 1977), “syntactic clitics” vs. “phonetic clitics” (Zaliznjak 2008: 8), “root-like clitics” vs. “affix-like clitics” (Aikhenvald 2002), “2P clitics” vs. “head-adjacent clitics” (Bošković 2001; Peng and Billings 2006; Franks 2008), “clustering clitics” vs. “non-clustering clitics” (Zimmerling and Kosta 2013), “clause-level clitics” vs. “phrase-level clitics” (Spencer and Luís 2012; Zimmerling 2013: 71), “ditropic clitics” (Cysouw 2005), “exoclitics” vs. “endoclitics” (Harris 2002), etc. The book’s terminology is standard, apart from the terms “diacclisis” and “pseudodiacclisis” introduced in Chapter 2 and illustrated in §8.10. They describe word orders in which clustering clitics linked with the same clausal head (*diacclisis*) or with different clausal heads (*pseudodiacclisis*) do not assume a contact position; cf. example (23) on p. 34: *po gradovima* =su¹ *predsednici opština* =se¹ *odjednom opredeljivali*¹ ‘in the cities, the municipality presidents were suddenly deciding...’ (Bosnian). A regular trigger of pseudodiacclisis in BCS is clitic climbing, when the climbed clitic does not reach the position of clusterization in the higher clause. This option is discussed at length in the second part of the reviewed book (cf. pp. 227, 266–73, 292, 307–9, 318–20, 371). Meanwhile, true clause-bound diacclisis is apparently a more marginal phenomenon in BCS (p. 168). A straightforward explanation of this asymmetry is that in a class of the world’s languages including BCS, contact position of the same clause’s clustering clitics is the default principle, while in configurations with clitic climbing, the contact position of the matrix- and embedded-clause clitics is just a tendency; although the matrix-clause clitic template has slots for all classes of the climbed elements, i.e., for BCS or

⁶ In contrast to Franks and King 2000 and the subsequent tradition in Slavic studies, the authors of the reviewed book do not use the label “AUX” for verbal clitics, since they see no difference in the ordering of copular and auxiliary forms of BCS *biti* ‘be’ (p. 19). I believe this is merely a question of the “depth” of analysis in terms of Haspelmath 2019, i.e., the linguist’s readiness to implement the apparatus of formal theories: clitic templates can be analyzed both as generalizations over text data and as ordering algorithms. In most Slavic languages, the constraints on placement of auxiliary present-tense forms of the verb BE, especially in the 1st and 2nd person, are more rigid compared to the copular uses of BE. Therefore, the uses of auxiliary present-tense BE-clitics in the Slavic perfect provide diagnostic contexts for the identification of clusters. In some Slavic languages, there is no marked contrast between auxiliary and copular present-tense BE-clitics; some languages also extend this analogy to other types of auxiliaries. That means that the AUX slots diagnosed by the auxiliary uses of present-tense BE-clitics can attract other clustering elements representing more recent layers of Slavic clitics, cf. the BCS future auxiliary *htjeti*.

Czech reflexives and pronominal argument clitics, the extracted clitics do not always reach these slots by movement and are sometimes left in intermediate positions.

A linguist needs valid research tools and resources like language corpora in order to measure the variation and, last but not least, have models that can be evaluated on text collections. These issues are discussed in Chapter 3, where the authors argue for a “triangulation of methods” using the scheme INTUITION/THEORY—OBSERVATION—EXPERIMENT and state that many theoretical claims concerning BCS clitics remain controversial and have not been checked properly against empirical data (p. 56). Most observations made in the reviewed book are based on existing BCS corpora; cf. Chapter 8 for Bosnian, Chapter 13 for Serbian, Chapter 14 for Croatian, and general preliminaries for corpus analysis in chapters 4 and 12. An experiment conducted with 336 Croatian speakers is presented in Chapter 15. The design of this experiment is somewhere in between socio- and psycholinguistics. On the one hand, the authors test a set of stimuli with and without clitic climbing and get the acceptability judgments of the experimental subjects. On the other hand, they measure the subjects’ reaction time post-operationally: as Figure 15.5. (p. 383) shows, the processed reaction time is in the range from 1,000 to 8,000 ms, i.e., from one to eight seconds. In this situation, both cover terms—“sociolinguistics” and “psycholinguistics”—are possible. The authors opt for the latter (pp. xii, 16, 49, 53, 57, 59), which is not surprising given their stance on not aligning their project with sociolinguistics⁷. Chapter 4 concludes the opening third of the book entitled “Preliminaries” and serves as a brief guide to the existing Bosnian, Croatian, and Serbian corpora providing data about the regional varieties of BCS.

The main part of the book is divided into two parts. Part 2 (chapters 5–9), entitled “Parameters of Variation”, offers a detailed discussion of most parameters, except for those related to clitic climbing. The latter are discussed in Part 3, “Clitic Climbing” (chapters 10–15). This subdivision has two motivations. First, as the authors argue, most systemic variation, i.e., variation within one and the same idiom of BCS, is attested in configurations licensing clitic climbing. Second, the descriptions of clitic climbing in Slavic languages bring in additional syntactic factors, such as the distinction of raising vs. control predicates, internal structure of finite and non-finite embedded complements of a different type, clause restructuring, etc. These factors are less relevant for the parameters of clitic ordering discussed in Part 2. Chapter 17 provides a general summary to parts 2 and 3. Appendices A and B contain details about the design of the stimuli used in Chapter 15 and explain the statistical measures used in chapters 14 and 15.

⁷ The term “sociolinguistic” is used in the book 22 times, the term “psycholinguistic” 38 times.

Turning back to the second part of the book, I would like to comment on three chapters. It was a wise move to dedicate a special chapter (Chapter 6) to linguistic traditions behind the descriptions of BCS clitics. There is little variation in the inventory of clustering clitics, but only one linguist—the Bosnian author Ridjanović (2012: 440)—openly claims that the reflexive marker *se* (historically an accusative form of the Proto-Slavic reflexive pronoun) lacks case in Modern BCS, since the parallel dative form *si*, which is widely used in Croatian, can hardly be found elsewhere in BCS territory (p. 99). This echoes the situation in Old Russian, where the accusative reflexive form *sja* is a highly frequent clitic, while its dative counterpart *si* occurs rarely and is not characteristic for vernacular Old Russian texts close to oral speech (Zaliznjak 1993: 284; 2008: 35). However, Zaliznjak puts the reflexive clitics *sja* and *si* in the same slots, ACC and DAT, that host Old Russian argument clitics. His decision is motivated by the fact that there are no Old Russian examples where *sja* and non-reflexive accusative pronouns (cf. *mja*.1SG.ACC, *tja*.2SG.ACC, *i*.3SG.ACC.M, *ju*.3SG.ACC.F, etc.) combine,⁸ and the same restriction holds for ORus *si* and non-reflexive dative pronouns (cf. *mi*.1SG.DAT, *ti*.2SG.DAT⁹). To assess the claim that BCS *se* and BCS/Croatian *si* lack morphological case, I would like to get more genuine examples where *se* and BCS accusative clitics, Croatian *si*, and dative clitics occur as parts of the same cluster.¹⁰ Such sentences are scattered elsewhere in the book, but it would be helpful to put them together. Other issues with templatic orders are the ordering of accusative and genitive clitics¹¹ and the dropping of the auxiliary *je*.AUX.3SG in the sequence *se je*. On

⁸ For modern Slavic languages, the recognition of REFL as a separate template slot different from ACC is based on two facts: (i) the template order for the argument pronominal clitics in most Slavic idioms is DAT ACC, while the reflexive marker generally precedes dative clitics in the cluster with the order REFL [_{CL.ARG} DAT ACC]; and (ii) the accusative and dative reflexive clitics are generally ordered the same way.

⁹ Unlike modern South and West Slavic languages, Proto-Slavic and Old East Slavic dialects lacked 3rd-person dative clitics. The same gap is found in Old Church Slavonic and the oldest texts written in Old South Slavic idioms.

¹⁰ Browne (1975/2004: 257) points out that *se* and same-clause accusative clitics combine in reflexive impersonal sentences like *Veterani su uvidjeli, da =ih =se vara* 'The veterans realized that people were fooling them', accepted by some Croatian grammarians. This construction is mentioned by the authors of the reviewed book on p. 45, where they provide a colloquial Croatian parallel *Čuje*.3SG =SE.REFL *kišu*.ACC 'One hears the rain' to standard BCS *Čuje*.3SG =SE.REFL *kiša*.NOM. However, they do not give examples like *Čuje*.3SG =je.3SG.F =SE.REFL 'One hears it', where the accusative argument of the impersonal reflexive verb is realized by a clitic.

¹¹ The sets of BCS accusative and genitive clitics are almost identical, with the exception of 3SG.F forms, where the accusative clitic, but not the genitive one, has the allomorph *ju* (Browne and Alt 2004: 33). However, according to the authors of the reviewed book, who sum up the recommendations of Bosnian, Croatian, and Serbian

p. 100 the reader learns that there is a disagreement between two groups of BCS authors. The authors from the first group (cf. Piper and Klajn 2014: 452; Milićević 2007: 105) postulate two slots with the order ACC > GEN and accept sentences like *Lišili*.PTCP.3PL =SU.AUX.3PL =ih.3PL.ACC =je.3SG.F 'They_i deprived them_j of it' as standard. The authors from the second group (cf. Mrazović and Vukadinović 2009: 659; Ridjanović 2012: 565) deny that accusative and genitive clitics combine in standard BCS,¹² which means that there is only one ACC/GEN slot in their BCS idioms.¹³ Regarding the haplology rule *se je* > *se* called "haplology of unlikes" (pp. 104–05), the reported facts suggest that it is not a local phonetic modification, but a syntactic process. The default overt form of the 3rd-person singular perfect auxiliary is replaced by the zero auxiliary *je*.AUX.3SG > Ø.3SG; according to Ridjanović (2012: 564), standard Bosnian always retains the *se je* sequence in sentences like *Dobro*.PRED =se.REFL =je.3SG *nadati*.INF 'It is good to hope', where =*je* is not a perfect auxiliary, but a copula.¹⁴ The rest of Chapter 6 is devoted to two aspects of clitic-external ordering in BCS—first-word vs. first-constituent variation and delayed clitic placement¹⁵, when the clitics skip the initial constituent. Remarkably, the authors leave very short comments about delayed placement (p. 113), despite it being a more complicated mechanism that involves restructuring of the whole clause. This is probably due to the authors' plan to minimize the apparatus before the corpus study in Chapter 8, which is aimed at measuring the proportion of 2P and delayed clitic placement in the corpus of spoken Bosnian. Regarding normative descriptions of BCS, the authors state (p. 123) that Serbian linguists generally understand 2P, i.e., the default position of the clustering clitics, as the position posterior to the first spelled-out phrase (XP CL), while Croatian and Bosnian linguists generally understand it as the position posterior to the

grammarians on pp. 94–95, this morph disappears from many BCS idioms. The nouns have non-homonymic forms of ACC and GEN, except for the singular forms of masculine animate nouns.

¹² As far as one can tell from the reviewed book, this split of judgments is not between the regional BCS varieties, but between two modes of description: Serbian authors are found in both competing lines of analysis.

¹³ Clustering clitics *x*, *y* are put in the same template slot, if they meet two conditions: (i) they do not combine in a cluster, and (ii) they are ordered the same way regarding all other clustering clitics *u*, *w*.

¹⁴ The asymmetric behavior of the 3rd-person auxiliaries vs. 3rd-person BE-copulas is attested both in Modern and in Old Slavic languages. For example, Old Russian/Old East Slavic grammaticalized the zero 3rd-person perfect auxiliary in all number forms but retained overt 3rd-person copulas with nominal predicates in the same group of vernacular texts (Zaliznjak 2008: 259).

¹⁵ The abbreviation "DP = delayed position", introduced earlier on p. 28, is not optimal because of the association with the term "Determiner Phrase" (DP).

first phonetic word (2W) and favor the configurations with phrase splitting (X_1 CL $X_2 \dots X_n$) and delayed placement (X/XP Y CL).

Chapter 7, entitled “Clitics in Dialects (Bosnian, Croatian, Serbian)” (pp. 127–71), is a welcome complement to normative descriptions of BCS. The information rendered here can be broadly classified into two unequal groups: (i) the usage in Štokavian dialects close to regional BCS varieties, where the parameter settings for clitics (e.g., phrase-splitting and delayed placement conditions, haplology rules, etc.) can, with few adjustments, be derived from the settings of BCS grammar; and (ii) autonomous clitic systems incompatible with BCS. A large majority of facts pattern with the first group, with the exception of 1P orders in the absolute clause-initial position ($\#X$ CL...)—cf. examples (75–82) on pp. 160–62—and the dialects with endoclititics discussed on pp. 163–64.¹⁶ There are reasons to assume that endoclitisis, i.e., a configuration where clitics are inserted into morphological structure, is not an inherent feature of clitics, but a feature of certain *clitic bases* hosting the clitics. No languages where any elements are invariably realized as endoclititics are attested: in all known cases, endoclitisis is a side effect of some proclitics or enclitics attaching to special clitic bases, combining the features of syntactic and morphological structures. These criteria apply to splitting of superlative adjectives like *nâj* ↓ *drăžī* ‘dearest’,¹⁷ which can be realized in BCS dialects with endoclititics as *nâj* ↓ =*mi*.1SG.DAT =*je*.AUX.3SG *drăžī* ‘He is my dearest’; cf. examples (84–86). One more possible endoclititic basis is the future auxiliary *htjeti*: here, the ‘yes-no’ clitic marker *li* is inserted between the stem and the inflection; cf. example (89), reproduced below in a slightly modified notation:

- (1) Će ↓ =*li* =š jütre rivät tò storīt? (Čakavian)
 FUT FOC 2SG tomorrow manage.INF that get.done.INF
 ‘Will you be able to do it tomorrow?’

It is better to exclude examples (87) and (88), since the negator *ne* is traditionally considered a syntactic element, especially if it assumes a distant position from the verb.

Chapter 8 presents data retrieved from the corpus of spoken Bosnian. It is large enough to contain around 3,400 single clitics and 430 clusters, but nevertheless too undersized to provide statistics on clusters consisting of three elements or more. That means that in order to get all combinatorics of BCS

¹⁶ The authors do not specify whether BCS dialects with *vP*-internal clitics of the Romance type, where the distant position of clitics and verbs is banned ($*V$ X CL, $*CL$ X V) and the clitics lack a fixed position with respect to the clausal left boundary (“V-systems” according to Zimmerling and Kosta (2013), “verb-adjacent clitics” in traditional notation), are attested in the Štokavian area shown on p. 129.

¹⁷ The arrow ↓ marks the locus of endoclititic insertion.

clitics, one needs a much larger corpus. The combination *se*.REFL + *je*.AUX.3SG occurs in the Bosnian corpus only six times, with the order $=se$.REFL $=je$.AUX.3SG predicted by the BCS template, while the reverse order, $=je$.AUX.3SG $=se$.REFL, is attested 25 times. The haplological variant *se*.REFL + *je*.AUX.3SG > $=se$ Ø.3SG, where the auxiliary is deleted, occurs around 80 times (pp. 193–95). These figures are too small to establish whether the variation $=se$.REFL $=je$.AUX.3SG \sim $=je$.AUX.3SG $=se$.REFL represents non-identical template rules by different speakers or the lack of rigid ordering in the regional Bosnian variety of BCS. However, they indirectly confirm that the spelled-out combination *se* + *je* is undesirable, though not completely blocked in BCS. The authors measure the external position of Bosnian clustering clitics and find no difference in the placement of single clitics and clusters: around 94–95% of them end up in 2P, and in 77% of clauses, the clitics are placed after the first word (2W). The rate of delayed placement labeled “3P” is 4% (181 clauses). Twenty-six clauses (1%) pattern with the category “1P”¹⁸. These are not true instances of clause-initial clitics, but examples with clitics placed after parenthetical insertions; cf. example (55) on p. 211, reproduced below in a slightly modified notation:

- (2) Jedan drug #Musliman # $=me$ $=je$ (BCS, Bosnian)
 one friend Muslim me.1SG.ACC AUX.3SG
 zvaó...
 called.PTCP.SG.M
 ‘One friend, a Muslim, called me...’

From the perspective of clausal structure, examples like (2) are a special case of 2P with parentheticals intervening between clitic hosts ([_{NP/DP} *jedan drug*]) and clitics/clusters.¹⁹ However, the tag “3P”, referring to delayed placement in languages like BCS, can be taken at face value, since there is no obvious way to claim that the initial phrase is extraclausal. If the initial element has an effect on the end position of 2P clitics, shifting them to the right, it is a *barrier* in terms of Zaliznjak 2008 (esp. p. 48) and Zimmerling and Kosta 2013. Preliminary observations on diverse languages with clitics indicate that clitic orders

¹⁸ In Reinkowski 2001 (esp. p. 191), the term “initial position” (Ger *Anfangstellung*) is used differently; it refers to the placement of BCS clustering clitics after the first phonetic word. The authors of the reviewed book justly identify this option as a special case of 2P.

¹⁹ An analysis like this proceeds from the assumption that parentheticals split the already generated well-formed combinations of the 2P enclitics and their host category: $X/XP = CL$, $\# Y\# \Rightarrow X/XP \# Y\# = CL$ (cf. a similar approach in Franks 2017: 189–93). Some languages license post-syntactic insertions of this kind, others do not, but no language with clustering clitics licenses parenthetical insertions inside the cluster (Zimmerling 2013: 303–05).

with barriers have information-structural triggers, so that delayed placement configurations in BCS can be associated with some marked types of information structure.²⁰ The authors of the reviewed book do not delve into the issues of communicative-syntactic interface; they measure the heaviness of clause-initial elements and actual clitic hosts in clauses with 2P and 3P by the number of segments (graphemes). This measurement has been implemented in corpus studies of Old Czech clitics in Kosek et al. 2018. With 2P, the most frequent clitic host in spoken Bosnian is just two “graphemes” long,²¹ while with 3P, the initial element is three graphemes long, and the actual host, four graphemes long (p. 202). The deviations in the data are caused by rare over-long initial constituents ($n > 20$ graphemes). The authors exclude them from the sample, calculate the Wilcoxon signed-rank coefficient, and arrive at the conclusion that delayed placement in spoken Bosnian results from significantly long initial constituents that block 2P placement (p. 204). As far as I see, the procedure applied is correct, but additional data is needed to interpret the correlation between the length of constituents and 3P in linguistic terms. First, one needs a larger sample, where the length of initial constituents in both groups (2P and 3P) can be measured based on the number of words, not the number of graphemes. Second, one must try the alternative hypothesis that short initial constituents consisting of one or two words can trigger 3P, if they have some special communicative value.

The last third of the book opens with a brief introduction to the theory of clitic climbing (Chapter 10) and an extended comparison of clitic climbing phenomena in two Slavic languages: Czech and BCS (Chapter 11). The contrastive perspective is explained by the fact that the conditions for clitic climbing in Czech are better studied as compared to BCS (see the important works of Junghanns 2002; Rezac 2005; Hana 2007; and Rosen 2014 for Czech; and Stjepanović 2004 and Aljović 2005 for BCS), while the morphosyntax of both languages is similar. The necessary, but not sufficient condition for clitic climbing in languages like BCS or Czech is that the clitic template of the matrix/higher clause has slots for categories represented by the clitics extracted from embedded clauses.²² There are three groups of factors that

²⁰ Cf. the syntactic approach to clitic-third orders in Croatian (Ćavar and Wilder 1999) and an equivalent analysis of Croatian idioms in terms of barriers in Zimmerling 2013 (esp. pp. 454–63).

²¹ This result depends on the chosen transcription. For processing, the authors use the phonetic tags, like /De/ instead of the normalized spelling *gdje* (p. 187).

²² Standard word-order systems with 2P clitics (“W-systems” in the notation of Zimmerling and Kosta 2013 and Zimmerling 2013) manifest identical sets of clustering clitics in root and embedded clitics, but clitic climbing is theoretically possible even if these sets are non-identical. The same holds true for word-order systems of the Bulgarian–Philippine type, where the 2P condition is combined with clitic-verb

Franks and King (2000: 245) and Stjepanović (2004) argue that clitic climbing is only possible with clause restructuring when the predicative complement lacks the full properties of an autonomous clause. This is a viable alternative to the assumption that the complement always forms a clause on its own (Spencer and Luís 2012). Indeed, if a clitic cluster is a real syntactic object and not just a sequence of phonetically adjacent weak-stress elements, the clustering clitics must obey the One-Domain-Principle (Zimmerling 2021: 483), although it is a priori not clear whether this domain corresponds to a single clause or to a clause union. The authors of the reviewed book seem to adopt the restructuring hypothesis and try to prove it empirically. In Chapter 12, they introduce the design of the corpus studies and specify the details for retrieving and processing embedded finite *da*-clauses containing clitics. It is customary in Balkan studies to distinguish the uses of the particle *da* in indicative and subjunctive clauses (cf. Joseph 1983; Stjepanović 2004; Todorović 2015; Mitkovska, Bužarovska, and Ivanova 2017), but the authors stick to synonymous labels *da*₁ (for tensed indicative *da*-clauses) vs. *da*₂ (for tense-less subjunctive *da*-clauses) that go back to Browne 1968, 1986, 2003 (esp. p. 39). Since they focus on clitic climbing out of finite tensed complement clauses into tensed matrix clauses, they exclude matrix predicates with *da*₁-complements from the sample (p. 292). They also exclude two unwanted types of *da*₂-predicates: reflexive and polyfunctional. The list of complement-taking predicates (CTPs) checked in the remaining types of BCS *da*₂-clauses includes 17 items: the authors deleted the most frequent raising verbs from the list in order to make it more balanced, since object control verbs have a much lower frequency (ibid.). The corpus study of clitic climbing out of *da*₂-clauses in Chapter 13 is based on a Serbian corpus. The authors conclude that Serbian *da*₂-clauses marginally allow clitic climbing in raising and subject control contexts, but it is probably blocked with object control. There is no evidence that the reflexive *se* can climb out of *da*₂-clauses in any context. Clitic climbing out of tensed *da*₂-clauses is a marginal construction in Serbian, but the future/past-tense markers, contrary to previous claims, do not block it completely (pp. 308–09).

The impact of the raising vs. control asymmetry for clitic climbing out of infinitival complements is studied in Chapter 14 on Croatian data. In order to check these issues, one needs to take those BCS varieties where the use of infinitival complements is a living phenomenon: the distinction of Standard and Colloquial Croatian represented by different corpora adds an extra dimension. The list of verbs tested in this chapter is a bit larger and includes 24 items. The results show that clitic climbing occurs more frequently in Standard Croatian, where the difference between raising and simple control verbs is statistically significant, while clitic climbing with reflexive subject control CTPs is significantly less frequent in all registers of Croatian. The type of infinitive clitic and its case are not relevant (p. 325). Chapter 15 again

deals with the regional Croatian variety of BCS and the same input data—the infinitival complements with clitics—but the method is different. The authors test the set of stimuli on 336 Croatian speakers (non-linguists, students of Zagreb University, with the average age of 21.5 years). The set includes 40 verbs: 8 raising verbs, e.g., *moći* ‘can’, *trebati* ‘have to’, *počinjati* ‘start’; 8 non-reflexive subject control verbs, e.g., *znati* ‘know’, *uspijevati* ‘succeed’; 8 non-reflexive object control verbs, e.g., *pomagati* ‘help’, *dozvoljavati* ‘allow’; 8 reflexive object control verbs with the marker *si*, e.g., *braniti si* ‘forbid oneself’, *dozvoljavati si* ‘allow oneself’; and 8 reflexive object verbs with the marker *se*, e.g., *učiti se* ‘teach oneself’, *spremati se* ‘prepare oneself’ (pp. 335–38). At the output, the authors get two types of data—acceptability judgments and reaction time. The data of both types are processed by the same regression measure as in Chapter 14 (p. 315, 325, 354). The most important presented empirical result is that they got statistical confirmation that clitic climbing is not obligatory in BCS with any type of predicate, including raising verbs (\approx “restructuring predicates”), although the speakers had marked preferences for the clitic climbing order in this group of CTPs (p. 384). Morphological case of the infinitive clitic is relevant. If the controller is in the dative and the infinitival clitic is in the accusative, clitic climbing is possible, but the acceptability rate is still under 50%. The object control reflexive constraint, first postulated by Hana (2007), proved relevant to BCS, apart from the so-called lexical reflexives²⁴ (*bojati se* ‘be afraid’, *vratiti se* ‘return’, etc.), where climbing is marginally possible (p. 386). Chapter 16 concludes Part 3. The authors state that the conditions licensing clitic climbing are heterogeneous and argue that their interaction, as well as the optionality of many rules, is a case of systemic complexity in the spirit of Rescher 1998 and Miestamo, Sinnemäki, and Karlsson 2008.

The title of the reviewed book asserts that it is about “clitics in the wild”, i.e., clitics as they are. One could say that it is a successful attempt at *taming* the clitics and making a bridge between theoretical models and empirically oriented linguistics. Turning to complexity, I would like to add two dimensions. The first one is the perspective of (mathematical) formal grammars. Clitic ordering apparently corresponds to three classes of them. Clause-bound cluster orders (in a different terminology, template orders) can be generated by A(utomaton)-grammars based on the immediate predecessor relation; they produce the string stepwise from the edge slot to the adjacent one and look up one single template slot per step, but if the ordering is rigid, this algorithm works. The 2P placement principle can arguably be modeled by a broader class of grammars—context-free grammars—given the reliable assumption that clitics cluster in dedicated syntactic positions; in languages like BCS, Czech, Pashto, Ossetic, Warlpiri, etc., it is 2P (X/XP — CL). The X ~ XP variation in 1P licensed in BCS or Warlpiri, but not in Czech or Ossetic, is a

²⁴ This label is first introduced on pp. 48–49.

special form of the 2P condition and a parametric setting characteristic of a subclass of 2P languages. However, the orders with delayed placement (\approx 3P) and clitic climbing can hardly be generated by context-free grammars. The main reason is not that the rules of delayed placement and clitic climbing are non-obligatory, but that they involve reordering of the already generated structure due to such context factors as topicality of the initial phrase or some active lexical or grammatical feature, etc. This reordering can only be done by context-sensitive grammars, or at best, mildly context-sensitive grammars with movement operators (Stabler 1997, 1998; Gärtner and Michaelis 2007). Mildly context-sensitive grammars generate the structure bottom-up but capture both right-to-left movement, i.e., raising, and left-to-right movement, e.g., lowering (cf. Zimmerling 2021: 431).²⁵ The second dimension is interface phenomena. Although both information layering and syntactic derivation are complex processes, the principles of the communicative-syntactic interface must be simple; otherwise, the speakers would not be able to apply them. It is tempting to assume that reordered clauses with clitics are also communicatively marked. This is likely for delayed clitic placement, since initial barriers in languages with fixed-position clitics are generally topical, i.e., add a fixed information-structural value (cf. sentences (3a) and (4a) above), but less evident for BCS clitic climbing data; the status of neutral vs. communicatively marked word orders has to be established for different groups of complement-taking predicates with embedded-clause clitics on a separate basis. The overall ratio of the default and marked orders and its dynamic are important; they show whether a word-order system is diachronically stable or not.²⁶ According to the reviewed book, the ratio of delayed placement order in the corpus of spoken Bosnian totals only 4–5% (p. 197), while the historical study of Reinkowski (2001: 182, 201), mentioned on p. 114, shows that delayed placement orders are “dominant” in the corpus of Serbian and Croatian journalistic texts from 1903 up to 1995. These discrepancies of data can be explained by at least three different factors: (i) the oral vs. written contrast; (ii) the level of text complexity—the journalistic texts are likely more complex and provide more opportunities to use initial topical barriers than the fragments included in the spoken corpus—and (iii) the size of the corpus.

²⁵ The unilateral restriction on the movement vector does not change the efficiency of Stablerian mildly context-sensitive grammars. The Minimalist program, as is well known, accepts only right-to-left movement.

²⁶ An increase in frequency of the marked order can lead to its reanalysis as the new default setting. This happened, for example, to Old Russian accusative reflexive *sja*, which was a 2P clitic in the 11th to 12th centuries but ended up in the postverbal position in the 15th to 16th centuries due to barrier rules shifting it to the right (Zaliznjak 2008: 169–220).

I conclude that *Clitics in the Wild* is a valuable contribution to Slavic studies. This book gives the reader what has been promised—an analysis of microvariation in BCS clitic syntax—and serves as a modern guide to a number of issues in general grammar. I am happy to recommend the book to all Slavicists and other linguists.

References

- Aikhenvald, Alexandra. (2002) "Typological parameters for the study of clitics, with special references to Tariana". R. M. V. Dixon and Alexandra Y. Aikhenvald, eds. *Word: A cross-linguistic typology*. Cambridge: Cambridge University Press, 42–78.
- Aljović, Nadira. (2005) "On clitic climbing in Bosnian/Croatian/Serbian". Nedžad Leko, ed. *Lingvistički vidici*. Sarajevo: Međunarodni forum Bosna, 58–84. [Forum Bosnae, 34.]
- Billings, Loren. (2004) "Review of *On the nature of the syntax-phonology interface: Cliticization and related phenomena*". *Journal of Slavic linguistics* 12(1–2): 285–321.
- Bošković, Željko. (2001) *On the nature of the syntax-phonology interface: Cliticization and related phenomena*. Amsterdam: Elsevier.
- Browne, Wayles. (1968) "Srpskohrvatske enklitike i teorija transformacione gramatike" [Serbo-Croatian enclitics and transformational grammar theory]. *Zbornik za filologiju i lingvistiku* [Zbornik matice srpske za filologiju i lingvistiku] 11: 25–29.
- . (1975/2004) "Serbo-Croatian enclitics for English-speaking learners". Rudolf Filipović, ed. *Contrastive analysis of English and Serbo-Croatian/Kontrastivna analiza engleskog i hrvatskog ili srpskog jezika*. Vol. 1. Zagreb: Institute of Linguistics, 105–34. [Reprinted in *Journal of Slavic linguistics* 12(1–2): 249–83.]
- . (1986) *Relative clauses in Serbo-Croatian in comparison with English*. Zagreb: Institute of Linguistics. [Yugoslav Serbo-Croatian-English Contrastive Project: New Studies, 4.]
- . (2003) "Razlike u redu riječi u zavisnoj rečenici: Kontaktni i distantni položaj veznika *da*₂ i glagola/Unterschiede der Wortstellung im abhängigen Satz: Kontakt- und Distanstellung der Konjunktion *da*₂ zum Verb" [Differences in word order in dependent clauses: Contact and distance position of the conjunction *da*₂ and the verb]. *Wiener Slawistischer Almanach* 57: 39–44. Available at: <https://www.slavistik.uni-muenchen.de/forschung/publikation/almanach/sonderbaende/almans57.htm>.
- Brown[e], Wayles and Theresa Alt. (2004) *A handbook of Bosnian, Croatian and Serbian*. Durham, NC: Slavic and East European Language Research Center (SEELRC), Duke University. Available at: <http://www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=1>.

- Ćavar, Damir and Chris Wilder. (1999) "'Clitic third' in Croatian". Henk van Riemsdijk, ed. *Clitics in the languages of Europe*. Berlin: Mouton de Gruyter, 429–67. [Empirical Approaches to Language Typology, 5.]
- Coseriu, Eugenio. (1980) "'Historische Sprache' und 'Dialekt'" ["Historic language" and "dialect"]. Joachim Göschel, Pavle Ivić, and Kurt Kehr, eds. *Dialekt und Dialektologie: Ergebnisse des Internationalen Symposiums "Zur Theorie des Dialekts," Marburg/Lahn, 5–10. September 1977* [Dialect and dialectology: Proceedings of the International Symposium "On the theory of dialect," Marburg/Lahn, 5–10 September 1977]. Wiesbaden: Steiner, 106–22. [Zeitschrift für Dialektologie und Linguistik: Beihefte, 26.]
- Cysouw, Michael. (2005) "Morphology in the wrong place: A survey of the proposed enclitics". Wolfgang U. Dressler, Dieter Kastovsky, Oskar E. Pfeiffer, Franz Rainer, eds. *Morphology and its demarcations: Selected papers from the 11th Morphology Meeting, Vienna, February 2004*. Amsterdam/Philadelphia: John Benjamins, 41–104. [Current Issues in Linguistic Theory, 264.]
- Franks, Steven. (2008) "Clitic placement, prosody and the Bulgarian verbal complex". *Journal of Slavic linguistics* 16(1): 91–137.
- (2017) *Syntax and Spell-Out in Slavic*. Bloomington, IN: Slavica.
- Franks, Steven and Tracy Holloway King. (2000) *A handbook of Slavic clitics*. New York/Oxford: Oxford University Press.
- Gärtner, Hans-Martin and Jens Michaelis. (2007) "Some remarks on the locality conditions and minimalist grammars". Uli Sauerland and Hans-Martin Gärtner, eds. *Interfaces + recursion = language? Chomsky's minimalism and the view from syntax-semantics*. Berlin: De Gruyter, 162–95. [Studies in Generative Grammar, 89.]
- Hana, Jiřka. (2007) *Czech clitics in higher order grammar*. Ph.D. dissertation, Charles University in Prague.
- Harris, Alice. (2002) *Endoclititics and the origins of Udi morphosyntax*. Oxford: Oxford University Press.
- Haspelmath, Martin. (2019) "Ergativity and depth of analysis". *Rhema* (2019) 4: 108–30. DOI 10.31862/2500-2953-2019-4-108-130.
- Ivanova, Elena, and Anton Zimmerling. (2019) "Shared by all speakers? Dative predicatives in Bulgarian and Russian". *Bulgarian language and literature* (2019) 4: 353–63.
- Joseph, Brian D. (1983) *The synchrony and diachrony of the Balkan infinitive: A study in areal, general, and historical linguistics*. Cambridge: Cambridge University Press.

- Junghanns, Uwe. (2002) "Clitic climbing im Tschechischen" [Clitic climbing in Czech]. Uwe Junghanns, ed. *Untersuchungen zur Syntax und Informationsstruktur slavischer Deklarativsätze* [Studies on the syntax and information structure of Slavic declarative clauses]. Leipzig: Institut für Linguistik der Universität Leipzig, 57–90. [Linguistische Arbeitsberichte, 80.]
- Kosek, Pavel, Radek Čech, and Olga Navrátilová. (2018) "Starobylá dativní enklitika *mi, si, ti* ve staročeské bibli 1. redakce" [Ancient dative enclitics *mi, si, ti* in the 1st redaction of the Old Czech Bible]. Petr Malčík, ed. *Vesper Slavicus: Sborník k nedožitým devadesátinám prof. Radoslava Večerky*. Prague: Nakladatelství Lidové noviny, 137–51.
- Miestamo, Matti, Kaius Sinnemäki, and Fred Karlsson, eds. (2008) *Language complexity: Typology, contact, change*. Amsterdam/Philadelphia: John Benjamins Publishing Company. [Studies in Language Companion Series, 94.]
- Miličević, Jasmina. (2007) "Co-occurrence of Serbian second-position clitics: Syntactic and morphonological constraints". Nabil Nathout and Fabio Montermini, eds. *Morphologie à Toulouse: Actes du colloque international de morphologie 4èmes Décembrettes* [Morphology in Toulouse: Proceedings of the international conference on morphology Décembrettes 4]. Munich: Lincom Europa, 99–121. [Lincom Studies in Theoretical Linguistics, 37.]
- Mitkovska, Liliana, Eleni Bužarovska, and Elena Ju. Ivanova. (2017) "Apprehensive-epistemic *da*-constructions in Balkan Slavic". *Slověne* 2: 57–83.
- Mrazović, Pavica and Zora Vukadinović. (2009) *Gramatika srpskog jezika za strance* [Grammar of Serbian language for foreigners]. 2nd revised edition. Sremski Karlovci: Izdavačka kuća Zorana Stojanovića.
- Peng, Adam and Loren Billings. (2006) "Binukid pronominal clisis". Paper presented at the 10th International Conference on Austronesian Linguistics, 17–20 January 2006, Puerto Princesa City, Palawan, Phillipines. Available at: <http://www.sil.org/asia/Philippines/ical/papers.html>. Last accessed 20 February 2020.
- Piper, Predrag and Ivan Klajn. (2014) *Normativna gramatika srpskog jezika* [Normative grammar of Serbian]. 2nd revised and amended edition. Novi Sad: Matica srpska.
- Reinkowski, Ljiljana. (2001) *Syntaktischer Wandel im Kroatischen am Beispiel der Enklitika* [Syntactic change in Croatian exemplified by the enclitic]. Munich: Otto Sagner. [Slavistische Beiträge, 405.]
- Rescher, Nicholas. (1998) *Complexity: A philosophical overview*. New Brunswick, NJ: Transaction Publishers.
- Rezac, Milan. (2005) "The syntax of clitic climbing in Czech". Lorie Heggie and Francisco Ordóñez, eds. *Clitics and affix combinations: Theoretical perspectives*. Amsterdam: John Benjamins Publishing Company, 103–40. [Linguistik Aktuell/Linguistics Today, 74.]

- Ridjanović, Midhat. (2012) *Bosnian for foreigners: With a comprehensive grammar*. Sarajevo: Rabic Publishing Company.
- Rosen, Alexander. (2014) "Hapology of reflexive clitics in Czech". Elżbieta Kaczmarek and Motoki Nomachi, eds. *Slavic and German in contact: Studies from areal and contrastive linguistics*. Hokkaido: Slavic Research Center, 97–116. [Slavic Eurasian Studies, 26.]
- Spencer, Andrew and Ana R. Luís. (2012) *Clitics: An introduction*. Cambridge: Cambridge University Press.
- Stabler, Edward. (1997) "Derivational minimalism". Christian Retoré, ed. *Logical Aspects of Computational Linguistics: First International Conference, LACL '96, Nancy, France, September 23–25, 1996. Selected papers*. Berlin: Springer, 68–95. [Lecture Notes in Artificial Intelligence, 1328.]
- Stabler, Edward. (1998) "Acquiring languages with movement". *Syntax* 1(1): 72–97.
- Stjepanović, Sandra. (2004) "Clitic climbing and restructuring with 'finite clause' and infinitive complements". *Journal of Slavic linguistics* 12(1–2): 173–212.
- Todorović, Nataša. (2015) *The indicative and subjunctive da-complements in Serbian: A syntactic-semantic approach*. Frankfurt am Main: Peter Lang. [Potsdamer Linguistische Untersuchungen, 16.]
- Zaliznjak, Andrej A. (1993) "K izučeniju jazyka berestjanyx gramot" [Toward the study of the language of the birchbark letters]. Valentin L. Janin and Andrej A. Zaliznjak, *Novgorodskie gramoty na bereste: Iz raskopok 1984–1989 godov* [Novgorod birchbark letters: From the excavations of 1984–1989]. Moscow: Nauka, 191–319.
- (2008) *Drevnerusskie enklitiki* [Old Russian enclitics]. Moscow: Jazyki slavjanskoj kul'tury.
- Zimmerling, Anton V. (2013) *Sistemy porjadka slov slavjanskix jazykov v tipologičeskom aspekte* [Word ordering systems of Slavic languages in the typological aspect]. Moscow: Jazyki slavjanskoj kul'tury.
- (2018) "Dva dialekta russkoj grammatiki: Korpusnye grammatiki i model'" [Two dialects of Russian grammar: Corpus grammars and model]. Vladimir P. Selegey et al., eds. *Computational linguistics and intellectual technologies: Papers from the Annual International Conference "Dialogue"* (2018): Issue 17(24). Moscow: RGGU, 818–30. Available at: https://www.dialog-21.ru/media/5217/_dialog2018scopus.pdf.
- (2021) *Ot integral'nogo k aspektivnomu* [From integral to aspective]. Moscow/St. Petersburg: Nestor-Istoria.
- (2022) "Historical text corpora and the conclusiveness of linguistic analysis". Vladimir P. Selegey et al., eds. *Computational linguistics and intellectual technologies: Papers from the Annual International Conference "Dialogue"* (2022): Issue 21. Moscow: RGGU, 586–93. Available at: https://www.dialog-21.ru/media/5847/_dialog2022scopus.pdf.

- Zimmerling, Anton and Peter Kosta. (2013) "Slavic clitics. A typology." *Sprachtypologie und Universalienforschung (STUF)* 66(2): 178–214.
- Zwicky, Arnold. (1977) *On clitics*. Bloomington, IN: Indiana University Linguistics Club.

Anton Zimmerling
Pushkin State Russian Language Institute
Institute of Linguistics, Russian Academy of Science
Moscow, Russia
fagraey64@hotmail.com