Acta Linguistica Petropolitana. 2020. Vol. 16.2. P. 306–328
DOI 10.30842/alp2306573716211

The Balkan sprachbund in typological-geographical space[*]

J. Nichols
University of California, Berkeley (USA); Higher School of Economics, Moscow (Russia); University of Helsinki (Finland); johanna@berkeley.edu

Abstract. The Balkan sprachbund is known to its specialists as distinctive and made up of languages that are closely similar, a view that can best be assessed by typological comparison. This paper compares three different western Eurasian sprachbunds — Balkan, Circum-Baltic, and the Avar sphere in the Caucasus — to each other and to the larger sets of western Eurasia, all of northern Eurasia, and all of the northern hemisphere. The typological properties compared are six complex typological macrofeatures each consisting of a set of related features. They capture some of the classic Balkan features and some deep-seated typological traits, and they enable us to place the Balkan area in the larger typological and linguistic-geographical map of Eurasia. Mapped in typological space, the Balkan languages prove to be discrete as a set and to form a compact cluster at or even beyond the edge of the European typological space. That is, the Balkan sprachbund occupies an apex position, hyper-European or even ultra-European, in a typological map of Eurasia. In these respects it differs from the Circum-Baltic area and in some respects resembles the Avar sphere, which is also very compact and often peripheral or extreme in the Caucasus. The Balkan-Avar similarities and differences can be accounted for by similarities and differences in the sociolinguistics of the two areas. The main general conclusion is that the Balkan sprachbund is essential to understanding the linguistic geography and typology of all of Europe. Issues for future research are testing the robustness of the hyper-European characterization by surveying more features, and determining whether the Balkan sprachbund is leading the evolution of a European linguistic profile or is a zone of peripheral archaisms in an evolutionary process trending in the opposite direction.

Keywords: sprachbund, language area, Balkan, Circum-Baltic, Avar sphere, linguistic typology, causative, causal-noncausal pair, inflectional person category, noun-based language, verb-based language, event structure, linguistic complexity, finite verb.

[*] Supported in part by a grant from the Russian Academic Excellence Project 5-100 to the Higher School of Economics, Moscow; and a Helsinki University Humanities Visiting Professorship, 2017–2019. I thank Andrew Dombrowski for carrying out much of the Balkan data collection.

Балканский языковой союз в типологическом и географическом пространстве

Дж. Николс

Университет Калифорнии, Беркли (США); Высшая школа экономики, Москва (Россия); Университет Хельсинки (Финляндия); johanna@berkeley.edu

Аннотация. Балканский языковой союз известен соответствующим специалистам как особый тип языковой общности, состоящей из языков, обнаруживающих значительное сходство. Эта точка зрения может быть наилучшим образом проверена с помощью типологического сравнения. В данной статье три западноевразийских языковых союза — балканский, циркумбалтийский и аварский на Кавказе — сравниваются между собой и с более обширными выборками языков Западной Евразии, всей Северной Евразии и всего северного полушария. Сравниваемыми типологическими характеристиками являются шесть сложных типологических макропризнаков, каждый из которых включает в себя набор связанных признаков. Они охватывают как некоторые классические балканизмы, так и глубинные типологические особенности и позволяют нам локализовать балканский ареал на более обширной типологической и лингвогеографической карте Евразии. В типологическом пространстве балканские языки занимают особое место как общность и образуют компактный кластер на периферии и даже за пределами европейского типологического пространства. Можно сказать, что балканский языковой союз на типологической карте Евразии занимает вершинную — «гиперевропейскую» или даже «ультраевропейскую» — позицию. В этом отношении от отличается от циркумбалтийского ареала, а в определенном плане напоминает аварский регион, который также очень компактен и занимает во многом периферийное положение в кавказском ареале. Балканско-аварские сходства и различия можно объяснить сходствами и различиями в социолингвистической ситуации в двух ареалах. Основной вывод заключается в том, что балканский языковой союз очень важен для понимания лингвогеографии и типологии всей Европы. Задачами для будущего исследования являются проверка надежности полученной гиперевропейской характеристики путем анализа большего количества признаков и выяснение того, возглавляет ли балканский языковой союз путь эволюции, характерный для языков европейского типа, или представляет собой зону периферийных архаизмов, сохранившихся в результате противоположно направленного эволюционного процесса.

Ключевые слова: языковой союз, языковой ареал, балканский, циркумбалтийский, аварский регион, лингвистическая типология, каузатив, каузативные и некаузативные глаголы, флективное выражение категории лица, язык именного типа, язык глагольного типа, лингвистическая сложность, финитный глагол.

1. Introduction

Typological properties are distributed non-randomly in space, and they figure in definitions of language areas and macroareas, the Balkan sprachbund being probably the first linguistic area to be so defined on the basis of structural features. This article addresses four questions about the linguistic geography of the Balkan sprachbund: 1) Where are the Balkan languages located in a typological map of Europe? A typological map plots structural features against each other and/or against geographical properties such as latitude-longitude coordinates, altitude, overland distance, etc. and draws typological, historical, and other conclusions from the distribution. Specifically for the Balkan languages, we want to know whether they are found at the edge or the center of the map, and thus whether they represent prototypical, extreme, or outlier members of the European and Eurasian linguistic populations. 2) In such a map, do the Balkan languages form a cluster? Areal linguistics focuses primarily on defining areas based on features that are unique to them or differ significantly in frequency between members and non-members of the area, and bona fide linguistic areas can be expected to show up as clusters in typological maps. 3) How do we define the diagnostic Balkan areal features to make them cross-linguistically surveyable? Here two commonly recognized Balkan features, loss of infinitives (or dispreference for their use) and clitic pronouns (especially those that can be doubled by independent arguments), are placed in a larger typological context. 4) What other typological features have distributions similar to those of the Balkan languages, and how do we interpret the distributions?

The overall message will be that the Balkan area is essential to understanding the linguistic geography of Europe. Typologically, it occupies an extreme position along each of several typological dimensions defining greater vs. lesser approximation to a European typological profile. As an extreme representative of an areal ideal, this can be called an **apex distribution** or **apex position** in a larger gradient. It is specifically **hyper-European** or **ultra-European** depending on whether it is at the edge or beyond it. The impetus for this approach comes from [Kortmann 1977: 225–231], where it is shown that, among European languages, the percentage of adverbial subordinators incorporating quantifying expressions like ‘how much’ (e.g. English *inasmuch as* ‘since’) is highest in Romance and Balkan languages, with Albanian at the highest position (e.g. Albanian *me sa* (with how:much) 'since, inasmuch as, insofar as'); and adverbial subordinators specifically of degree incorporating 'how much' are highest in Balkan languages and peak in Albanian.

In what follows, *Section 2* describes the design and method of the survey. *Section 3* describes the results: the distributions of the typological features over the areas. *Section 4* is discussion and conclusions.

2. Survey design

2.1. Survey

This study surveys six complex (or multivariate) typological features and their distribution across Eurasia and more generally. A complex feature is one that involves not just a single variable and its values (e.g. presence vs. absence of an inclusive/exclusive distinction) but a larger number of subvariables and their values. The features described in *Section 3* have about 20 to 50 subfeatures, each subfeature with two or more values. They are tracked across Balkan languages, nearby languages, and two other European areas, the Circum-Baltic area and the Avar sphere (also sometimes other languages of the Caucasus). All of these surveys are still ongoing, but while not complete they still make firm conclusions possible.

2.2. Areas: their sociolinguistics and known histories

The three principal areas surveyed are the Balkan sprachbund, the Avar sphere, and the Circum-Baltic area. The Balkan area and Avar sphere are a good minimal pair for comparison as they are similar in traceable age, size, widespread adult multilingualism, and many avenues for copying and convergence. Both formed in larger high-diversity regions (the Balkan peninsula, the Caucasus) but are sociolinguistically and typologically distinctive in those areas. The distinctive traditional sociolinguistics and convergence are no longer active, having ended in the early 20th century with the rise of universal education, national languages, and mass media, but for both areas we have descriptions reaching back to the active period and/or describing the usage and language competence of the oldest generations. (See [Dobrushina 2013] for reconstructed sociolinguistics of the late 19th–early 20th century using field interviews; also [Dobrushina et al. 2020]) All three are in accretion zones, regions where languages move in more often than out and diversity accumulates.

The Balkan area is not described in detail here as its history, sociolinguistics, and structural properties are presumed known to readers of this volume. The languages surveyed are standard Macedonian, standard Bulgarian, Greek, Torlak Serbian, Albanian, Arumanian, and Kalderaš Romani (data is often lacking for Torlak Serbian and Arumanian). (Also surveyed were near-Balkan languages around the Balkan area: Italian, BCS, Slovene, Romanian [standard], Italian, Turkish, Hungarian.) Language diversity and strong contact effects likely go back millennia, but the sprachbund as known to science probably took shape with the Ottoman conquest in the middle ages. The essentials of the traditional sociolinguistic situation include adult multilingualism, child monolingualism, situation-based language choice, limited code switching, discrete lexicons due to resistance to lexical borrowing, and partly convergent grammars where the convergent features are generally not native to any of the languages but often involve selection for analyticity [Friedman 2011; Joseph 2010; Lindstedt 2000, 2019; Lindstedt, Salmela in press]. Ethnic identity is fairly discrete and native language is part of it.

The Avar sphere (briefly described in [Nichols 2018]; fuller description underway; see also [Dobrushina et al. 2020]) approximately coincides with the area of the Avar khanate in the eastern Caucasus (the western part of today's republic of Daghestan). The Avar khanate arose in the 13th century when Avars took over the previous Sarir kingdom (5th–12th centuries CE), at which point Avar apparently began to spread at the expense of its close sister Andic languages. The khanate was a voluntary confederation of many small city-states each centering around a village or set of villages, each typically speaking its own language. The languages were all from the Nakh-Daghestanian (or East Caucasian) family: most languages of the Andic and Tsezic branches and most varieties of Avar. The Andic family is probably about 2000 years old, Tsezic somewhat older, and the split between the two still older; Avar is a sister language of all Andic, with considerable dialect divergence but reported mutual intelligibility of all Avar dialects. There is a good deal of adult bilingualism and multilingualism, varying from speaker to speaker, and Avar could generally be used as a lingua franca as it was widely known (since men served in the army where Avar was the language of command, and since the largest markets were in the Avar-speaking lowlands). The population was settled farmers and herders, but the male population was transhumant as most working-age males spent the winter half of the year away from their home villages in lowland winter pastures or urban centers where they held jobs or owned businesses. The region was characterized by asymmetrical vertical bilingualism whereby highlanders often learned lowland languages but not vice versa (since highlanders traveled downhill for markets and work). As a consequence, vocabulary and structural traits, dialects, and even languages tended to spread uphill. At least in the lowlands there appears to have been vacillating dominance of Avar and Andic speech until Andi lost its military superiority in the 17th century. There was minimal language and ethnic identity; identity lay mostly with clan, village, and religious organizations. Code switching was not resisted and appears to have been common. Lexical borrowing was common and there was considerable grammatical convergence with distinctly lower grammatical complexity than in other Nakh-Daghestanian-speaking areas, for the most part involving not analyticity (as in the Balkans) but transparency, regularity, and reduction of allomorphy in inflectional paradigms. There was no language mixture, but a situation of linguistic symbiosis where two (or more) languages could function together in a single discourse or utterance (via easily tolerated code switching) while remaining discrete overall. Attractors (targets of universal bias) such as rhyming and alliterating pronouns, light verb constructions, and causativization, diffused and expanded in the sphere.

The Circum-Baltic area [Dahl & Koptjevskaya-Tamm eds. 2001; Seržant in press] comprises the languages bordering the Baltic Sea, a maritime contact zone going back to at least the Viking era. There is a good deal of contact, most of it commercial and involving ports and coastal trade colonies, with Finnic and Saami the initial and main donor languages. Bilingualism patterns were local, mostly individual, varied, and in many cases probably occasional or ephemeral. Recent linguistic impact is mostly lexical, especially involving written and commercial language; older patterns are grammatical. At various times one or another language dominated commerce and trade, and trading centers and forts were established, but there has been no area-wide language spreading. The languages surveyed here are (in clockwise order) Kildin Saami, North Saami, Swedish, Norwegian, Danish, German, Lower Sorbian, Polish, Russian, Lithuanian, Latvian, Estonian, Finnish. (Near-coastal vernaculars, such as varieties of Low German, Baltic High German, Novgorod Old Russian, and Livonian would probably be more revealing survey objects, but the survey features are not accessibly covered for enough of them.)

2.3. Method

The six typological features are described in *Section 3* together with the findings about their distributions. The survey tracks the features across those three areas, Europe more generally, and/or all of Eurasia, and a set of near-Balkan languages (Slovene, Bosnian/Croatian/Serbian, standard Romanian, Italian, Turkish, Hungarian). Plots of feature values against other feature values, and feature values against longitude, are used to determine whether, and to what extent, the languages of the three areas form clusters, whether they overlap with each other typologically or are discrete, and to what extent they are typical of Europe or Eurasia.

3. Results

3.1. Causal-noncausal pairings

These are pairs of verbs such as *fear: scare* or *break* intrans. vs. trans.), or triads such as *sitzen : sich setzen : setzen*, all involving one or two noncausal forms (continuous *fear*, *sitzen*, bounded *sich setzen*) and a causal form (*scare*, *setzen*), where within each set the lexical semantics is the same and the verbs differ in continuous/bounded/causative. The typological issue is which of the verb forms is base for the other(s) and which is derived, typologized as the proportion of the pairs that is causativizing, decausativizing, etc. For this approach see [Nedyalkov 1969; Nichols 1982; Haspelmath 1993; Nichols et al. 2004; Grünthal & Nichols 2018 and work in preparation]. (1) shows some of the values of the variable.

(1) Some illustrative causal-noncausal pairs from a larger worldwide sample. The relevant morphology is bold. (Hyphens and clitic boundaries are not orthographic.)

	Non-causal	**Causal**	**Derivation**
	‘break’	**‘break something’**	**type**
Czech	*lomit=***se**	*lomit*	Decausative
Spanish	*romper=***se**	*romper*	Decausative
Aymara	*p’aki-***si***-*	*p’aki-***ña***-*	Double
Kazakh	*syn-u*	*syn-***dyr***-u*	Causative
German	*brech-en*	*brech-en*	Ambitransitive

The survey uses the 18 causal-noncausal pairs of [Nichols et al. 2004] (wordlist and instructions available as [Nichols 2017a]). The counts here use only the first nine pairs (the verbs with animate S/O), as the inanimates are less well attested in dictionaries.

Figure 1 shows the worldwide distribution of high vs. low proportions of causativization. (The southern continents are not fully sampled yet.) Self-evidently, causativization is rare in Europe and frequent elsewhere. *Figure 2* plots percent (of the nine verb pairs) causativizing x percent decausativizing for the Balkan and Circum-Baltic areas, and *Figure 3* shows just the Balkan area in the sample for all of Europe. They show that the Balkan languages are at the extreme low range of causativization (horizontal axis) and fairly high in decausativization, and form a compact group, with the exception of outlier Kalderaš Romani, which is a relatively recent arrival in Europe and apparently not yet Europeanized in this respect. Thus low causativization defines Europe (*Figure 1*) and the Balkan languages form a compact cluster at the extreme European range (*Figures 2–3*, p. 314); put differently, the Balkan sprachbund is a compact prototype of Europe in this respect. The Circum-Baltic area, in contrast, is very diffuse and ranges from high to low in both causativization and decausativization. The Avar sphere (*Figure 4*, p. 320) is fairly compact and at the high causativizing edge. (Note the many languages in the Caucasus with zero incidence of decausativization.) *Figure 5* (p. 320) compares the fit of the two areas to their larger contexts: the Balkan sprachbund is within the European population (the lines marking the standard deviations overlap considerably) and extreme for that population, while the Avar sphere is entirely outside of the rest of its family and the overall Caucasus population.

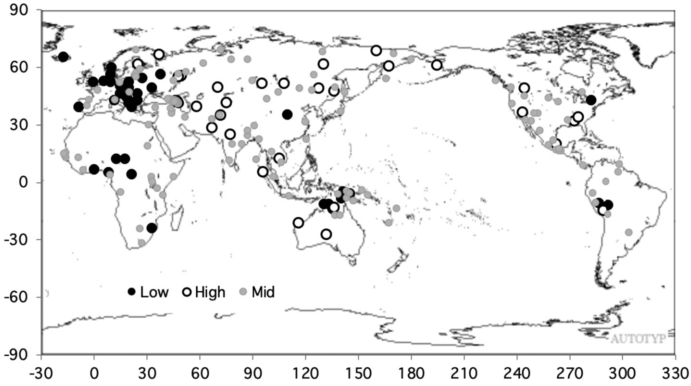

Figure 1. Languages with proportions 1 s. d. below (black) and above (white) the mean percent causativized. Gray = intermediate. ($N \sim 200$)

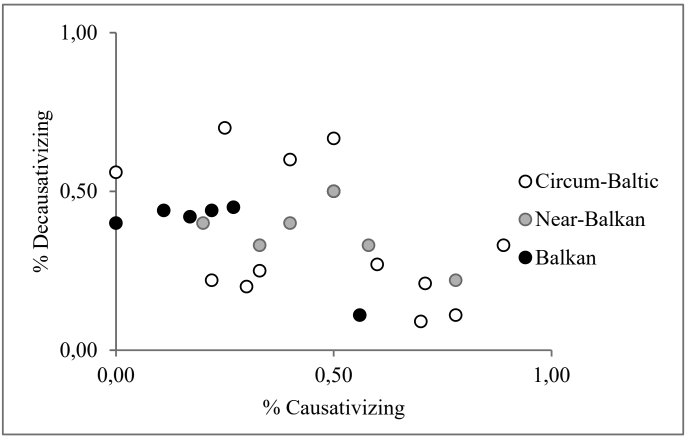

Figure 2. Percent decausativizing vs. percent causativizing: Balkan, Near-Balkan, and Circum-Baltic areas. Circles represent individual survey languages.

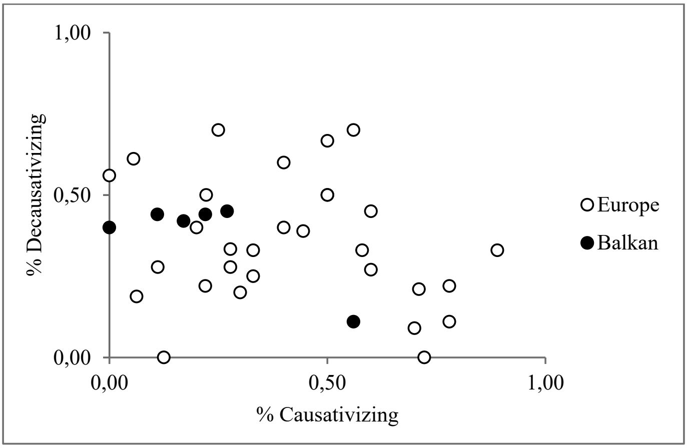

Figure 3. Percent decausativizing vs. percent causativizing: the Balkan area in Europe

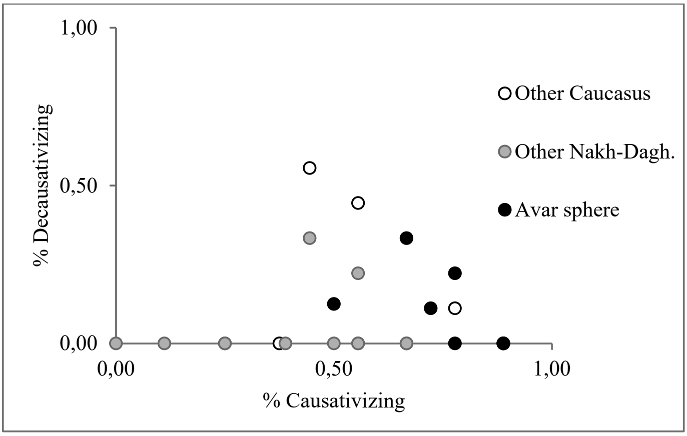

Figure 4. Percent decausativizing vs. percent causativizing in the Avar sphere

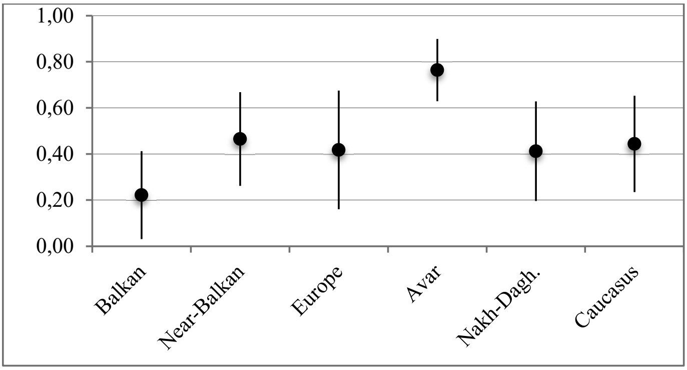

Figure 5. Mean percent causativized: Balkan sprachbund and Avar sphere in their larger areas

3.2. Lexical vs. inflectional person [1]

The category of person is borne by person markers (the term used in [Siewierska 2004, 2011]) such as independent pronouns, agreement affixes, and clitics and marked on other words (typically phrase heads) through

[1] This section is based on [Nichols 2017b].

agreement and other syntactic processes. In its various forms the person category may be more or less lexical or more or less inflectional. It is inflectional if it exhibits properties typical of inflection, such as agreement or pronominal argument marking on verbs or possessed nouns, appearance in outermost positions in inflected words, or formal marking in paradigms that share categories and/or structures with paradigms known to be inflectional. An example of the latter is the Turkish pronouns in (2).

(2) Turkish independent pronouns, singular and plural (nominative case):

1sg	**2sg**	**1pl**	**2pl**
ben	*sen*	*biz*	*siz*

The personal pronouns distinguish the same number categories as nouns and other pronouns, though *-iz* is not the regular plural ending and *-en* is not a singulative marker. Person behaves inflectionally in these respects in the Turkish paradigm, though of course it is also lexical as it is part of the basic meaning of each form.

Person is lexical if person markers have features of nouns such as inflecting in the same cases as nouns, using the same case morphology as nouns, exhibiting the same declension classes as nouns, or sharing other classification (e.g. gender) with nouns, and of course if it is inherent in lexemes (as it is in the Turkish pronouns of (2). The typology is based on the percent of items from a 50-item questionnaire that are inflectional or lexical. Here I track only inflectional person (the two are largely complementary, so either one gives much the same result).

Inflectional person displays the same kind of large-scale cline seen in most of the features surveyed here: extreme values appear in Europe (low in this case) and somewhere between North Asia and North America (high), with the cline especially clear in the higher latitudes of the Northern Hemisphere. *Figures 6* and *7* (p. 317) show the worldwide and northern higher-latitude distribution of inflectional person, with the percent inflectional person plotted against longitude[2]. Western Europe and west Africa are at the left and eastern America at right. There is a west-to-east upward trend in the worldwide plot (*Figure 6*), which is much steeper in the higher latitudes (*Figure 7*). There is a slight trend in the lower northern latitudes (not shown

[2] The trendline is for comparison only. It is calculated as a linear trendline, when in reality longitude lines are not parallel so correlations with longitude are not linear. Here and below, only the visible steepness is meaningfully comparable.

here; see [Nichols 2017b]) and next to none in the southern hemisphere. The Balkan languages are typically European. They are not at apex levels because they have increased their inflectional person marking by adding object agreement with person-marking clitics to the inherited subject agreement, but even with this innovation they are solidly within European values.

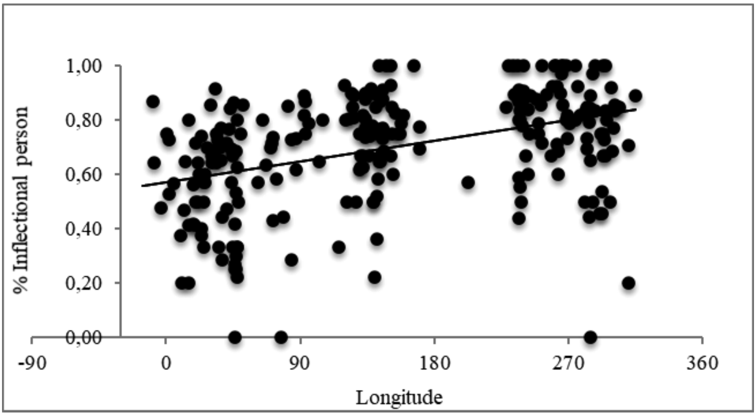

Figure 6. Inflectional person x longitude: World. ($N = 256$, $p <$ 000001)[3]

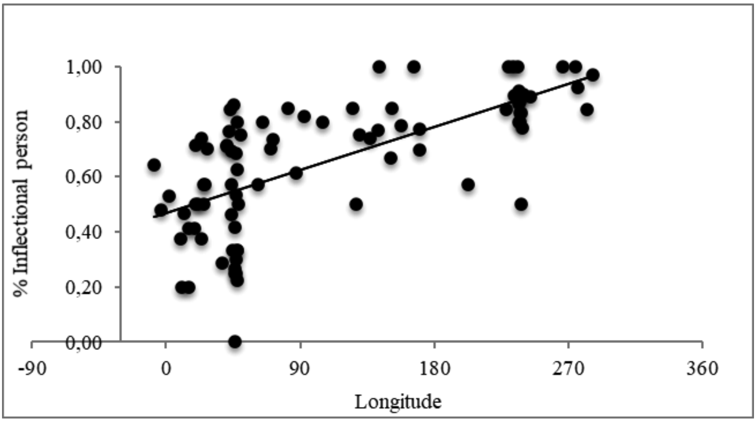

Figure 7. Inflectional person x longitude: Northern hemisphere above 40° N. ($n = 88$, $p < .000001$)

[3] Here and below, in plots against longitude significance is assessed with Spearman's rank correlation test.

Figures 8–9 show the distributions of the areas in the larger European contexts. The Balkan area is a tight cluster in the high range of inflectional person, with Arumanian a high outlier. The Circum-Baltic area, in contrast, is very wide-ranging and essentially identical to all of Europe. The Avar sphere stands out, in the Caucasus and in general, for its low inflectional person values. A distinctive trait of the languages of the Avar sphere is their lack of person indexation on verbs and on possessors; they have gender indexation on some verbs and no indexation on others. The Balkan area is fairly high, as are the Romance languages in general (not separately

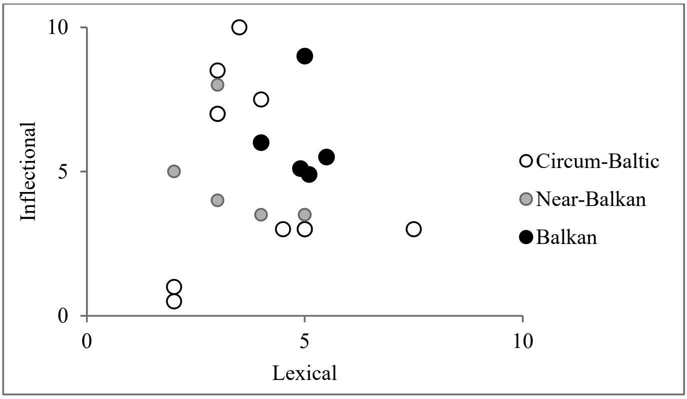

Figure 8. Number of lexical points x number of inflectional points:
The Balkan and Circum-Baltic areas in Europe. The Balkan high outlier is Arumanian

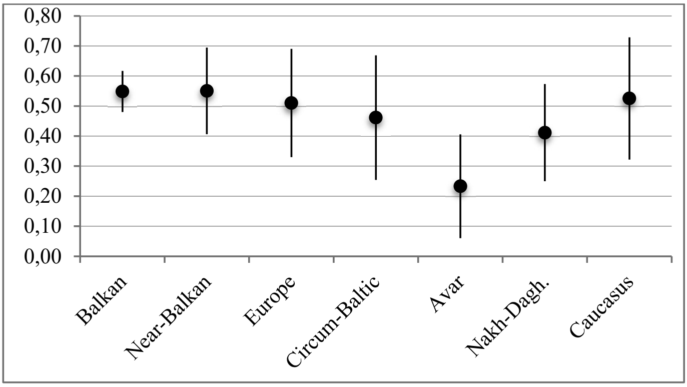

Figure 9. Mean percent inflectional: Three areas in context

identified in the graph), and it is no accident that the Balkan high outlier is Arumanian. The Balkan and Romance languages together are at the high edge of Europe, but this is not an apex distribution for Europe, where inflectional person is generally fairly low.

3.3. Noun-based vs. verb-based lexical type[4]

This feature concerns derivational sets like *fear : fright : frighten* or *white : whiten,* where one of the words represents the *base* which the others are derived from. In these examples *fear* and *white* are bases and the others derived from them by suffixation and (for *fear*, etc.) stem changes, chiefly vowel alternation. Late Proto-Slavic had the noun **strax-* (OCS *straxъ)* from which the verb *strax-i-* (OCS *strašiti*) was derived by suffixation. In German, however, the verb *fürcht-en* 'frighten' and the noun *Fürcht* 'fright, fear' are both basic (the verb ending *-en* is inflection and not part of the stem). In English the verb *sit* is basic and the noun *seat* derived from it by vowel change, and in Slavic **sěd-ě-*/ **sěd-* (OCS *sěděti, sěsti*) are basic and nouns such as modern Russian *siden'je* 'seat' are derived from the verb[5]. Thus individual paradigms differ in whether it is noun or verb (or adjective, adverb, etc., but noun and verb are most common) that is basic. Most languages have an overall favored type: in modern European languages nouns are more often basic and verbs more often derived, while in North America it is the reverse. The survey reported here used a wordlist of 50 such derivational sets, and the typology is based on the proportion that are verb-based, noun-based, adjective-based, flexible, etc. The survey is underway and the language coverage at this point is thin, but adequate to reveal a profile.

Here again there is a large west-to-east cline, with lower frequencies of the verb-based type in Europe and high frequencies in the Americas (*Figure 10*, p. 325), steeper in the northern hemisphere (*Figure 11*, p. 325).

Figure 12 (p. 326) shows the values for Europe. Of the three Balkan languages surveyed so far, Macedonian and Bulgarian are solidly European and Albanian is a conspicuous outlier with more verb-based sets. (This may

[4] This section is based on [Nichols 2016; Foley 2017], and unpublished work in progress.

[5] The *-ě- suffix of **sěd-ě-ti* is not derivation but a conjugation class marker, i.e. an extension.

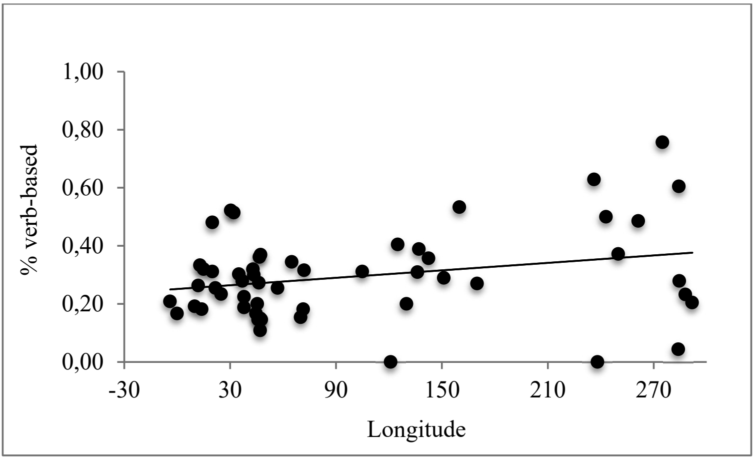

Figure 10. Proportion of verb-based sets plotted against longitude: Worldwide. (*N* = 67, not significant: *p* = 0.134)

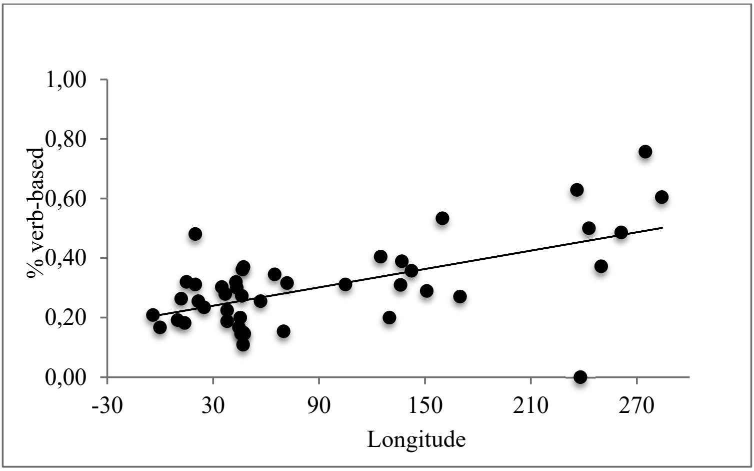

Figure 11. Proportion of verb-based sets plotted against longitude: Northern hemisphere. (*N* = 60, *p* = 0.012)

be an archaism in Albanian, since Proto-Indo-European was mostly verb-based, perhaps reflecting its origin farther east.) A major contributor to the European type is the high frequency of factitive morphology used to derive verbs from nouns in early Germanic and Slavic, as in Late Proto-Slavic **straši-i-ti* from **strax-ъ*[6]. Albanian evidently did not participate in this development. The Avar sphere is also not compact and is coextensive with the whole Caucasus. Overall it appears that both areas are typical of their larger regions, except for the outlier Albanian.

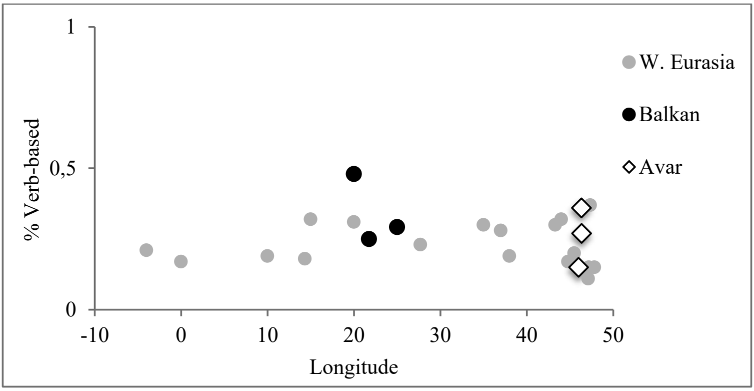

Figure 12. Proportion verb-based against longitude: Europe

3.4. Basic event structure

The distinction between verbs like Late Proto-Slavic **sěd-ě-ti* 'sit, be sitting' and ***sěd-ti* (> **sěsti*) 'sit down' will be termed *continuous* vs. *bounded* here, broad terms useful for cross-linguistic comparison where the distinctions in individual languages may be marked by language-specific actionality or aspect categories and may correspond to a variety of more specific event-structure categories such as (for continuous) state and activity, or (for bounded) telic, achievement, accomplishment, ingressive-stative, punctual, etc. Working at this broad level, for any continuous-bounded pair

[6] As is typical in such sets, the noun is itself deverbal, derived from **sterg- 'freeze, wary; guard' (see [Vasmer 1971/1987: 772] for the Balto-Slavic intransitive verb and [Derksen 2008: 467] for the transitive). But the ultimate deverbal origin is not part of the Late Proto-Slavic derivational set **strax- : straš-i-*.

one or the other verb (or both, or neither) may be basic in one or another language, and languages may have a clear preference for treating one or another as basic. This section is based on work underway and still in an early stage, using a wordlist of 24 verb pairs and 90 languages. (3) shows different choices for 'sit' verbs. In Spanish the two intransitive verbs are both derived, i.e. neither is basic, and the reason is that the base form is the causal, which is outside of the continuous-bounded pair.

(3) Continuous, bounded, and causal forms in selected languages. Base forms are bold.

	Continuous (state)	**Bounded (telic)**	**Causal**
English	***sit***	*sit down*	*seat*
German	***sitz***-*en*	*sich setz-en*	*setz-en*
Russian	*sid-e/i-*	***sed-***	*sad-i-*
Spanish	*est-ar sen-tad-o*	*sent-ar=se*	***sent***-*ar*

(4) shows results for posture verbs in languages from northern Eurasia and North America, calculating the proportion of verbs of that event structure type in the set of all posture verbs surveyed from that area. The bounded base type has its peak in northern Asia and is well represented everywhere. The continuous base type has its peak in North America. The causal base type, very rare worldwide, has its peak in western Eurasia, but the peak is a weak one: unlike the other peaks, this one is only within the column and not within the row, and it is outnumbered two to one by both of the other two types in the row (for base causals see again *§1*). Those are peak frequencies for the event structure base types (peaks within columns). Peak frequencies for continents (within rows) are bounded for western Eurasia (a somewhat weak preference, with continuous a close second) and northern Asia (strong preference), and continuous for North America (weak, with bounded a close second). The strongest preference shows up in northern Asia, where a low frequency for continuous cooccurs with a high frequency for bounded.

(4) Proportions of posture verbs with various bases.

Base:	**Continuous**	**Bounded**	**Causal**
Western Eurasia	0.36	0.45	0.16
Northern Asia	0.16	0.83	0.02
North America	0.54	0.46	0.00

Figure 13 shows the areal distributions in western Eurasia, plotting the number of base continuous verbs against longitude. Languages in the lower vertical range visibly increase to the right (= east), and in the Caucasus, a number of languages are piled on top of each other at the lower right. The Circum-Baltic languages are not clustered, though there is a local subcluster in the upper left (= west). The Balkan sprachbund is not clustered, though its mean is low; Macedonian is at the extreme European-like edge (the uppermost row of symbols) and the rest are in the lower ranges. The three that are closest to each other, in the lower part of the continuous range, are not in the expected range for extreme European values. The Avar sphere is at the low extreme, and hyper-compact, a cluster of superimposed symbols. Here the Balkan sprachbund is at the less European edge of the range, unless the position of Macedonian at the extreme European edge is an early move in the hyper-European direction (as would be consistent with the status of Macedonian as most Balkan of the Balkan languages). These findings are suggestive, though they could change as the study is completed.

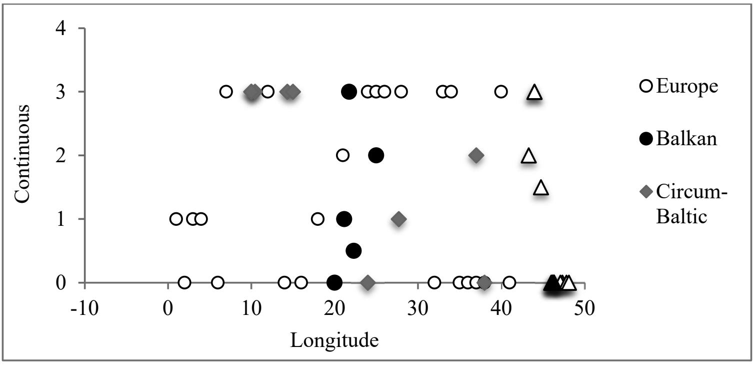

Figure 13. Numbers of posture verbs with continuous bases x longitude. A shadow under the symbol indicates two or more superimposed symbols.

3.5. Enumerative complexity

Enumerative complexity (a.k.a. taxonomic complexity, inventory complexity, and other terms) measures complexity as the number of items in the inventory for some domain. The measure used here (from [Nichols 2019, 2009]) counts the numbers of contrastive manners of consonant articulation, contrastive vowel qualities, tones, phonation types; syllable complexity (measured as the maximum number of consonants permitted in the syllable, whether at onset or coda); inflectional synthesis of the verb (number of categories marked on the verb: [Bickel & Nichols 2013]); and numbers of noun genders, classifiers, major or default alignments, and major or default word orders. The range for the languages covered so far in this survey runs from 9 to 27 enumerative complexity points.

Figure 14 plots the complexity levels against longitude for all four areas and the near-Balkan languages. The Balkan area is very compact, within Europe, and at an edge; the Circum-Baltic area is diffuse, extending the low-complexity range of Europe and reaching nearly to the highest range; and the Avar sphere is fairly compact and within the Caucasus. There is little overlap between Europe and the Caucasus. The complexity levels for western Eurasia are high compared to those of northern Asia, so the position of the Balkan languages at what appears in *Figure 14* to be the eastern range of western Eurasia is in fact still at the western edge of the overall range.

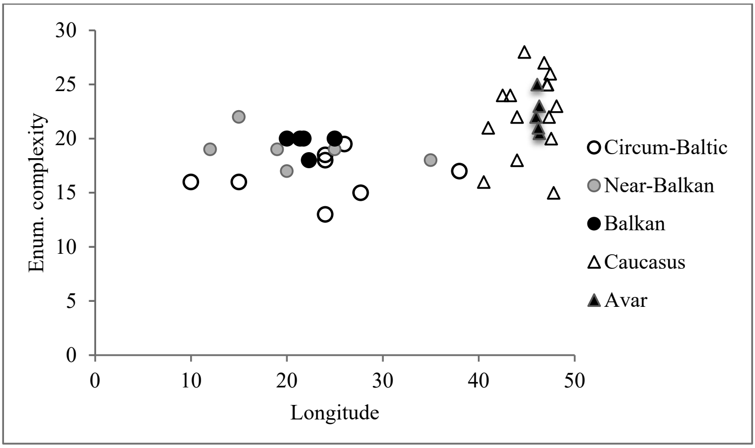

Figure 14. Enumerative complexity x longitude

3.6. Finiteness

[Shagal et al. 2019] survey whether verbs are finite or nonfinite in a number of exemplar constructions representing clause coordination, subordination, relativization, and complementation for various verb classes. They find a very steep gradient within Eurasia, where western languages use finite verbs in many more constructions than Siberian languages. The Balkan languages represent the extreme European range, with most of them using finites in all of the constructions. The classic Balkan trait of loss or minimization of infinitives is one specific manifestation of the tendency toward finiteness in all clauses. Defining the typological feature as finiteness makes it possible to compare the Balkan development with morphologically very different phenomena elsewhere (while loss of the Indo-European infinitive, as the Balkan feature is commonly described, is not amenable to comparison outside the Indo-European family). Defined this way, finiteness shows that the Balkan sprachbund is again a very compact area, consistent with the traditional view, and hyper-European, consistent with the general proposal here.

4. Discussion and conclusions

In summary, in most of the features surveyed here the Balkan sprachbund forms a compact cluster in typological or typological-geographical space, clear evidence of strong areality. It lies within the larger European population, but often at an edge, and where well enough sampled those edges tend to be hyper-European; in one case, finiteness, the Balkan languages lie beyond the edge and mark the apex not just of Europe but of all Eurasia. The Avar sphere is also compact with clear areality and often at the edge of the Caucasus or beyond the edge, usually on the far side from Europe. The Circum-Baltic area, in contrast, is typologically very diffuse, showing almost no areality, and mostly within Europe. Though more features need to be surveyed in order to characterize the areas accurately, even these first results are enough to show that the three areas differ from each other, in both their areality and the extent to which they are typical of their larger contexts.

The sociolinguistics of Balkan and Avar-sphere multilingualism is similar in that both involve adult multilingualism in a high-diversity larger population, but different in that the Balkan sociolinguistics keeps languages discrete and links them to identity, while in the Avar sphere there is code switching and other short-term mingling and language is minimally connected to identity. The structural consequences for the Balkan languages involve increased analyticity but no appreciable decomplexification; in fact mean complexity of Balkan languages is much higher than in Europe generally, and the standard deviations only barely overlap. This shows that decomplexification is not a necessary outcome of adult L2 learning and supports the claim of Lindstedt [2000, 2019] that analyticity, which makes morphemes easily identifiable and segmentable, can be more useful to L2 mastery than sheer non-complexity. The languages of the Avar sphere, in contrast, are decomplexified compared to their sisters[7], likely because easy code switching favors diffusion of selectively advantaged forms [Nichols 2018].

For purposes of cross-linguistic comparison, these various findings indicate that what is relevant to diachronic sociolinguistically-driven selection is not absolute feature values or yes/no distributions but the notions of peak vs. nonpeak or extreme vs. typical in a larger areal population of languages. What has not been shown is whether in its extreme placement the Balkan sprachbund is the cutting-edge leader in the evolution of a European linguistic profile, or a small cluster evolving locally. Either way, Balkan areal linguistics is essential to understanding the typology and evolutionary trends of all of Europe.

References

Bickel, Nichols 2013 — B. Bickel, J. Nichols. Inflectional synthesis of the verb. M. Haspelmath, M. Dryer, B. Comrie, D. Gil (eds.). *World atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. Available at: http://wals.info/chapter/22 (accessed Feb. 7, 2020).

Dahl, Koptjevskaja-Tamm 2001 — Ö. Dahl, M. Koptjevskaja-Tamm (eds.). Circum-Baltic languages. Amsterdam: Benjamins, 2001.

Derksen 2008 — R. Derksen. Etymological dictionary of the Slavic inherited lexicon. Brill: Leiden, 2008.

Dobrushina 2013 — N. Dobrushina. How to study multilingualism of the past: Investigating traditional contact situations in Daghestan. *Journal of Sociolinguistics*. 2013. Vol. 17. Iss. 3. P. 376–393.

Dobrushina et al. 2020 — N. Dobrushina, M. Daniel, Y. Koryakov. Languages and sociolinguistics of the Caucasus. M. Polinsky (ed.). *Oxford handbook of languages of the Caucasus.* Oxford: Oxford University Press, 2020. P. 163–192.

Foley 2017 — W. A. Foley. Structural and semantic dependencies in word class membership. N. Enfield (ed.). *Dependencies in language: On the causal ontology of linguistic systems*. Berlin: Language Science Press, 2017. P. 179–195.

[7] The difference is much stronger with the kind of complexity measure that considers transparency, consistency, and other biuniqueness properties [Nichols 2020].

Friedman 2011 — V. A. Friedman. The Balkan languages and Balkan linguistics. *Annual Review of Anthropology*. 2011. Vol. 40. P. 275–291.

Grünthal, Nichols 2018 — R. Grünthal, J. Nichols. Transitivizing/detransitivizing typology and language family history. *Lingua Posnaniensis*. 2018. Vol. 58. Iss. 2. P. 11–31.

Haspelmath 1993 — M. Haspelmath. More on the typology of inchoative/causative verb alternations. B. Comrie, M. Polinsky (eds.). *Causatives and transitivity*. Amsterdam: Benjamins, 1993. P. 87–120.

Joseph 2010 — B. Joseph. Language contact in the Balkans. R. Hickey (ed.). *The handbook of language contact*. Malden, MA: Wiley-Blackwell, 2010. P. 618–633.

Kortmann 1997 — B. Kortmann. Adverbial subordination: A typology and history of adverbial subordinators based on European languages. Berlin: Mouton de Gruyter, 1997.

Lindstedt 2000 — J. Lindstedt. Linguistic Balkanization: Contact-induced change by mutual reinforcement. D. G. Gilbers, J. Nerbonne, J. Schaeken (eds.). *Languages in contact*. Amsterdam; Atlanta: Rodopi, 2000. P. 231–246.

Lindstedt 2019 — J. Lindstedt. Diachronic regularities explaining the tendency toward explicit analytic marking in Balkan syntax. B. Joseph, I. Krapova (eds.). *Balkan syntax and (universal) principles of grammar*. Berlin: De Gruyter Mouton, 2019. P. 70–84.

Lindstedt, Salmela in press — J. Lindstedt, E. Salmela. Migrations and language shifts as components of the Slavic spread. T. Klír and V. Boček (eds.). *Language contact and the early Slavs*. Heidelberg: Winter, in press.

Nedyalkov 1969 — V. P. Nedyalkov. Nekotorye veroyatnostnye universalii v glagolnom slovoobrazovanii [Some probabilistic universals in verbal word formation]. F. Vardul (ed.). *Yazykovye universalii i lingvisticheskaya tipologiya* [Language universals and linguistic typology]. Moscow: Nauka, 1969. P. 106–114.

Nichols 1982 — J. Nichols. Ingush transitivization and detransitivization. *BLS*. 1982. Vol. 8. P. 445–462.

Nichols 2009 — J. Nichols. Linguistic complexity: A comprehensive definition and survey. G. Sampson, D. Gil, and P. Trudgill (eds.). *Language complexity as an evolving variable*. Oxford: Oxford University Press, 2009. P. 110–125.

Nichols 2016 — J. Nichols. Verb-based and noun-based languages. Presented at SLE 49, Leiden, 2016.

Nichols 2017a — J. Nichols. Realization of the causative alternation: Revised wordlist and examples. Manuscript. 2017. Available at: https://www.academia.edu/28861776/Causative_alternation_wordlist (accessed Feb. 7, 2020).

Nichols 2017b — J. Nichols. Person as an inflectional category. *Linguistic Typology*. 2017. Vol. 21. Iss. 3. P. 387–456.

Nichols 2018 — J. Nichols. Non-linguistic conditions for causativization as a linguistic attractor. *Frontiers in Psychology*. 2018. Vol. 8. P. 23–56.

Nichols 2019 — J. Nichols. Why is gender so complex? New typological considerations. F. Di Garbo and B. Wälchli (eds.). *Grammatical gender and linguistic complexity*. Berlin: Language Sciences Press, 2019. P. 63–92.

Nichols 2020 — J. Nichols. Canonical complexity. P. Arkadiev and F. Gardani (eds.). *The complexities of morphology*. Oxford: Oxford University Press, 2020. P. 27–66.

Nichols et al. 2004 — J. Nichols, D. A. Peterson, and J. Barnes. Transitivizing and detransitivizing languages. *Linguistic Typology*. 2004. Vol. 8. Iss. 2. P. 149–211.

Seržant in press — I. A. Seržant. The Circum-Baltic area. J. Fellerer and N. Bermel (eds.). *The Slavonic languages*. (Oxford guides to the world's languages). Oxford: Oxford University Press, in press.

Shagal et al. 2019 — K. Shagal, M. Wahlström, J. Nichols. (Non)finiteness in clause combining: A typological survey. Presented at 13th Conference of the Association for Linguistic Typology, University of Pavia. 2019.

Siewierska 2004 — A. Siewierska. Person. Cambridge: Cambridge University Press, 2004.

Siewierska 2011 — A. Siewierska. Person marking. J. J. Song (ed.). *Oxford handbook of linguistic typology*. Oxford: Oxford University Press, 2011. P. 322–345.

Vasmer 1971/1987 — M. Vasmer [M. Fasmer]. Etimologičeskiy slovar russkogo yazyka [Etymological dictionary of Russian]. Translated and annotated by O. N. Trubačev. Vol. 3. Reprint. Moscow: Progress, 1987. (First published in 1971).