Accepted for publication in *Studies in Language*

Distance-based approach reveals convergence effects in word order among the languages of the Circum-Baltic linguistic area

Ilja A. Seržant[1], Berfin Aktaş[2], Maria Ovsjannikova[1], Manfred Stede[3]
[1]- PoSla Typology Lab, Potsdam Slavic Variation Lab, Slavic Linguistics, University of Potsdam
[2] - Natural Language Understanding Lab, Technical University of Nuremberg
[3] - Applied Computational Linguistics, Department of Linguistics, University of Potsdam

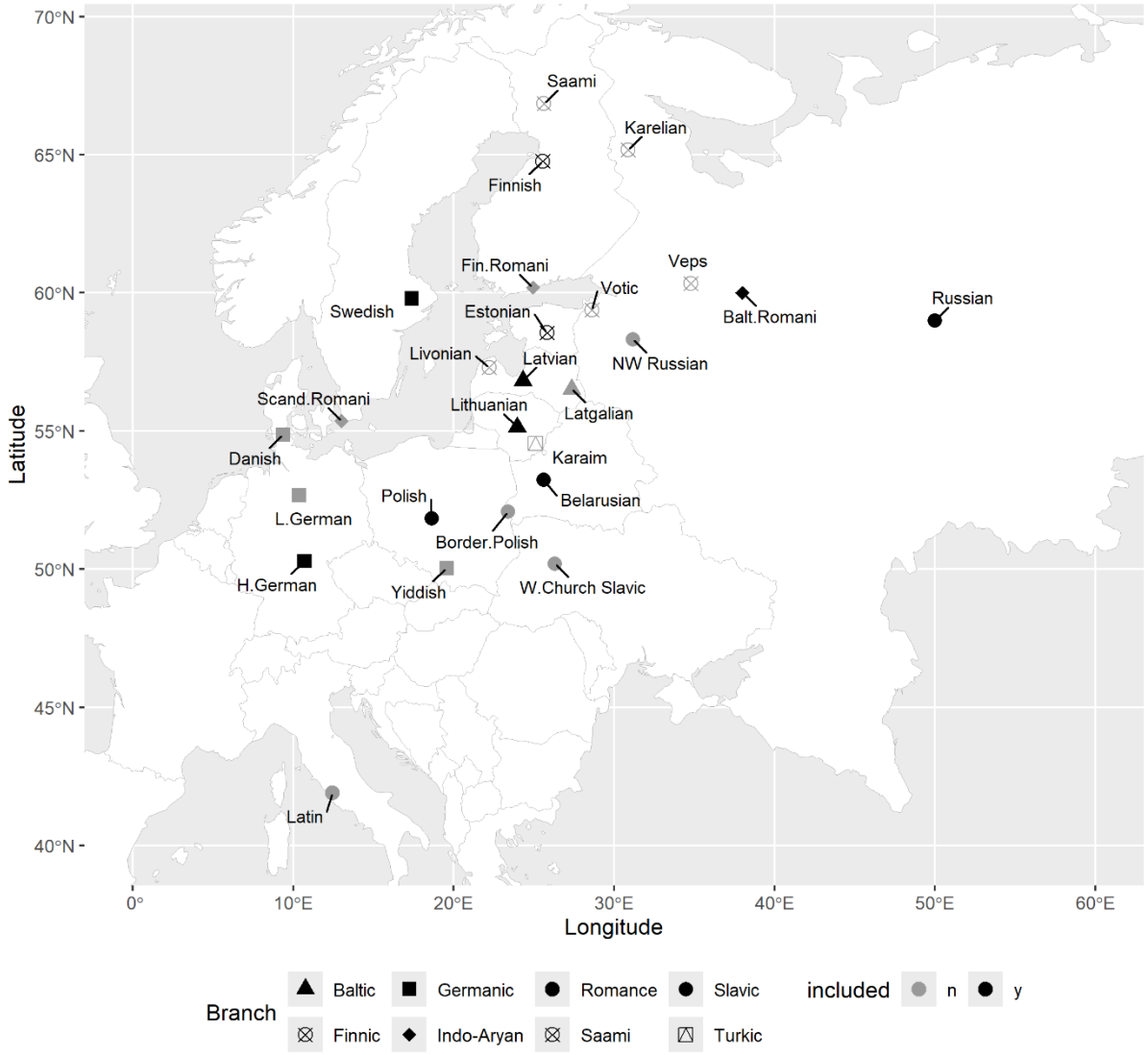
Abstract
We probe a new approach to linguistic areas that is based on positive dynamics towards an area rather than on full similarity of a feature across the languages of the area. Technically, we estimate the dynamics by comparing two distances: (i) the averaged pairwise distance between a language from the area (Focus language) and the other languages of the area with (ii) the averaged pairwise distance its genealogically closely related language outside of the area (Benchmark language) has to the languages of the area. This way we test whether the Focus language is closer to the area than its Benchmark language, which, if yes, we interpret as a trend towards convergence of the Focus language towards the languages of the area. We rely on Bible translations from 16 languages from the Parallel Bible Corpus (Mayer & Cysouw 2014) and examine the distances between the languages based on the order of words in the running text. We found that all languages of the CB area show convergence effects, with Baltic Romani and the two Baltic languages (Latvian and Lithuanian) being in the center of the area.

1. Introduction

The *Circum-Baltic (CB) area* – a term coined in Dahl & Koptjevskaja-Tamm (1992), perhaps after *baltischer Sprachbund* in Jakobson (1931[1971]: 137) (for other terms and subareas see Matthiassen 1985; Stolz 1991; Nau 1996), – is an established linguistic area along with the Balkan or Mesoamerican linguistic areas. Although there is no full consensus on which languages should belong to the CB area, the following languages are generally included: Indo-European languages: Polish, 'Borderland' Polish (*polszczyzna kresowa*) (West Slavic), Russian, North-Western Russian dialects, Belarusian, the West Russian variant of Church Slavic (East Slavic), Lithuanian, Latvian, Latgalian (East Baltic), Low German, High German, Yiddish (West Germanic), Swedish, Danish (North Germanic), marginally Latin (Romance) as well as the Baltic, Finnish and Scandinavian dialects of Romani (Indo-Aryan); it further includes most languages of the Finnic subfamily, such as Livonian (nearly extinct), Estonian, Finnish, Veps, Karelian, Votic, etc., and the Saami subfamily of the Uralic family. Finally, Karaim (Kipchak, Turkic) belongs here as well. The map in Figure 1 shows the geographical locations of the languages (the languages analyzed in the paper are rendered by black and the other by gray points).[1]

[1] The map was created in R (R Core Team 2024) using the packages *lingtypology* (Moroz 2017), *ggplot2* (Wickham 2016), *sf* (Pebesma & Bivand 2023; Pebesma 2018), *rnaturalearth* (Massicotte, South 2023), *rnaturalearthdata* (South, Michael, Massicotte 2024), and *ggrepel* (Slowikowski 2024). The coordinates for the languages, if available, were taken from Glottolog 5.1 (Hammarström et al. 2024), and in other cases assigned to the approximate centers of the spread of the lects in question, as indicated in the code, available at

Figure 1. Languages of the Circum-Baltic area



Historically, speakers of East Baltic, East Slavic and West Germanic languages immigrated into the coastal region of the Baltic Sea generally later than speakers of Finnic languages and assimilated some of them. Likewise, speakers of North Germanic immigrated into Scandinavia after speakers of Saami. Other Indo-European tribes (e.g., the now extinct branch of West Baltic) might have predated the arrival of the Finnic population in the area (see Kalio 2015; Lang 2018).

Methodological and conceptual problems such as defining the boundary of an area ("The boundary problem"), establishing the set of languages that should belong to the area ("The language problem") or establishing the set of features of an area in a non-arbitrary way ("The feature problem") (Masica 1976; Dedio et al. 2019: 499; van Gijn 2020; van Gijn & Wahlström 2023: 179-180) hold for the CB area too (see, *inter alia*, Nau 1996; Koptjevskaja-Tamm & Wälchli 2001).

REMOVEDFORANONYMYTY. The map indicates that High German as one of the languages analyzed in the paper, standing in this case for standard German.

The identification of this area crucially relies on a list of linguistic traits that are in one way or another similar across subsets of the languages of the area and are less or not at all characteristic of the surrounding languages not included in the area (see the overviews of such lists for the CB area in Koptjevskaja-Tamm & Wälchli 2001 or Seržant, aop). While such "list approach" provides a good approximation of what may single out the languages of a linguistic area against their broader geographical background, it has a number of limitations, which we discuss in detail in §2 below.

To overcome a set of methodological and theoretical shortcomings and limitations of traditional approaches, we suggest a new method that is crucially based on measurable distances between the languages within the area and related languages outside of it (henceforth the *distance-based approach*) with respect to a specific phenomenon.

We test our method on word order. Word order figures prominently in the discussion of the CB area in Wälchli & Koptjevskaja-Tamm (2001). They show that CB languages – very much like the Caucasus – represent a transitional area in the general European development from Eurasian SOV (Proto-Uralic, Proto-Turkic, Proto-Indo-European) to SVO and from genitive-noun to noun-genitive (Wälchli & Koptjevskaja-Tamm 2001: 709). However, they remain cautious about whether or not there are CB-specific areal effects on word order.

We proceed as follows. In Section 2, we first highlight some of the weaknesses of traditional approaches, which are crucially based on the similarity of the phenomenon in the languages of the area. In Section 3, we present our distance-based approach to linguistic areas. Section 4 gives a brief overview of word order patterns in the languages of the area. Section 5 presents our sample and the data. Section 6 discusses the specific computational methods used for the study to measure the distances. Finally, Section 7 presents the results of the study. In Section 8, we summarize and contextualize the results.

2. Why do we need a new approach to establishing convergence in linguistic areas?

It is thanks to traditional qualitative approaches that we know so much today about various linguistic areas around the globe. However, traditional approaches face quite a few limitations and methodological hurdles, which call for a new approach to estimating convergence in linguistic areas.

Areal linguistics emerged as an explanatory model of similarities across languages that is *complementary* to the explanations provided by the historical-comparative method, based on co-inheritance from the common ancestor language, as well as to the explanations based on universal preferences of languages (since Trubetzkoy 1923, 1928; Jakobson 1931). As a consequence, in areal research, the standard has been to rely solely on those convergent traits that are neither universally preferred (Haspelmath 2001: 1493) nor found in genealogically (closely) related languages. For this reason, linguistic areas are often required to consist of unrelated or distantly related languages (*inter alia*, Emeneau 1956: 124; Campbell 1985; Aikhenvald & Dixon 2001: 11; van Gijn 2020: 164).

However, there is no complementarity between genealogical, areal and universal factors (see recently Seržant, forthc.). Effects of language contact accumulate and lead to areal convergence in genealogically closely related languages too. Closely related languages may be even more prone to transfer and thus to convergence than unrelated or remotely related languages because of the greater similarity between the source and the target language prior to contact in the case of closely related languages (cf. the second factor in Matras 2007: 34). It has been argued that structural similarity of the languages in contact facilitates diffusibility of patterns (*inter alia,* Haig 2001; Epps et al. 2014 and critically Bowern 2014). It has also been shown that, along with geographical proximity, genealogy channels innovations (sound change across dialects in Heeringa & Nerbonne 2001).

Moreover, not only contact-induced innovations are likely to be found in closely related languages in contact. Shared inheritance itself also does not always exclude effects of language contact because such effects may also manifest themselves in a pressure to preserve some inherited traits, i.e., to contact-induced non-change (see, *inter alia*, Seržant 2021; Seržant et al. 2022). For example, Seržant (2021) shows that the preservation of the person-number inflection in Slavic languages from Proto-Indo-European was affected by language contact with the neighbouring Indo-European and non-Indo-European languages.

An extreme example of a linguistic area consisting of genealogically closely related lects is (dia)lectal areas (cf. van Gijn & Wahlström 2023: 185). Mutual contacts between (dia)lects is the reason why (dia)lects mostly do not drift apart but exchange innovations and exercise contact-induced pressures leading to the shared preservation of some of the inherited traits (see also Bowern 2013: 413-414). Thus, if our goal is to understand convergence in linguistic areas closely related languages should not be excluded.

Likewise, typologically frequent and universal phenomena may and do show areal skewing and may, therefore, be part of the descriptions of linguistic areas. For example, the two most frequent word orders – SOV and SVO – show areal skewing in Dryer (2013) such that SVO is typical for Europe and Southeast Asia while SOV for the rest of Eurasia (Dryer 2013). Universally preferred traits reflect preferences of human processor (or articulation apparatus) and, therefore, may even be more prone to borrowing and convergence than cross-linguistically rare or even dispreferred traits. Concededly, it is methodologically difficult to show that universally preferred traits are affected by language contact against the null hypothesis. For example, even though all languages of the Circum-Baltic area are SVO (Wälchli & Koptjevskaja-Tamm 2001: 704; Siewierska & Uhlířová 1998 on the Slavic languages), this cannot be regarded as a distinguishing feature of the languages of the area, since it is universally the second most frequent word order after SOV, see Dryer (2013). However, methodological difficulties should not a priori exclude universal traits from areal convergence.

Another problem with traditional approaches is their *selectiveness*. Since Trubetzkoy (1923, 1928), linguistic areas are traditionally described in terms of lists of areal traits (isoglosses or isopleths), see such lists for the CB area in Koptjevskaja-Tamm & Wälchi (2001) or Seržant (aop). Crucially, such traits are picked up by researchers primarily based on methodological considerations. The null hypothesis - i.e. no contact effects - can easily be ruled out with typologically rare traits because they are usually not found any close outside of the area. However, such lists may lead to skewed descriptions of linguistic areas, in which many potentially converging traits of the languages of the area remain unmentioned but, at the same time, infrequent and/or marginal traits are often included. For example, an established isogloss of the CB area is the presence of some phonemic tonal distinctions (Jakobson 1931; Koptjevskaja-Tamm & Wälchli 200: 640-646), compare the textbook example from Swedish *anden* 'duck.DEF' (pronounced with tone 1) vs. *anden* 'spirit.DEF' (pronounced with tone 2) or from Latvian *tā* (the falling tone) 'DEM.GEN.SG.M' vs. *tā* (the sustained tone) 'DEM.NOM.SG.F' vs. *tā* (broken tone with a glottal stop) 'so'. Since European languages outside of the area generally do not employ any tonemic distinctions (Maddieson 2013), it is likely that an areal impact must have promoted the retention (or even development) of at least some of the tonemic distinctions in CB. This isogloss is methodologically convenient because it is easy to prove as a property of the area against the null hypothesis. However, the informational contribution of tonemes in the languages of the area is minimal. There are only very few and only marginal minimal pairs and an L2 speaker not mastering distinct tones is never misunderstood. In this respect, the CB languages contrast to, say, languages of Southeast Asia, in which the distinctions of this type are crucial for successful communication.

In effect, the methodological rigor of the traditional approach to rule out the null hypothesis has the consequence that typological rara, cherry-picked and less salient phenomena become the best candidates for areal isoglosses, which, in effect, leads to inadequate descriptions of linguistic areas. As a consequence, considerable amount of structural parallelism among the languages of an area (see Civjan 1979; Bužarovska 2020: 59 for Balkan), sometimes referred to as mutual *translatability* (Gumperz & Wilson 1971: 154-155) or effects of *metatypy* (Ross 2007), cannot be explored and described in full despite the intuition that linguistic areas tend to converge towards one grammar, cf. "...roughly the same thing can be said in the same way…" (Campbell 2006: 4).

Finally, another methodological problem of traditional approaches to linguistic areas is that these are crucially *similarity-based*. That is, these approaches rely on assessed similarity of the phenomenon at issue across languages of the area (since Trubetzkoy 1928; see also Campbell 2006). Methodologically, however, without a clear baseline, similarity represents a serious problem since linguistic traits never exactly match. There will always remain some differences between similar phenomena across languages and, hence, some uncertainty and subjectivity as to whether or not the null hypothesis can be safely rejected. Differences between similar phenomena are not only found because language phenomena generally never exactly match but also for the following reasons.

First, convergent phenomena often arise via different historical pathways (this is the case with tonal distinctions discussed above), which unavoidably leads to somewhat divergent outcomes.

Second, histories of linguistic areas often consist of migrations and complex, layered contact configurations (Wälchli & Koptjevskaja-Tamm 2001). Since languages often migrate, hardly any linguistic area will attest language contact taking place over many thousands of years, achieving a high degree of convergence. For example, East Baltic and East Slavic languages occupied their current geographical locations in the CB area quite recently, no earlier than ca. 1000 years ago.[2] Yet, unless contacts last thousands of years it is likely that genealogically motivated traits will not only persist but will also dominate despite intensive contacts.

Third, languages normally do not arrive in an area all at once. The durations of pairwise contact may strongly vary across the languages of an area and, therefore, their degrees of convergence.

Fourth, the specific historical, political, social and environmental processes may constrain and skew the degree of convergence among subsets of languages in an area (*contact configuration* in Seržant 2021), as has been repeatedly emphasized in the literature (Nichols 1992; Tosco 2000; specifically for CB in Nau 1996; Wälchli & Koptjevskaja Tamm 2001). For example, Livonian speakers must have all been bilingual in Latvian and the contact effects of Latvian on Livonian were accordingly much stronger than the effects of, say, German on Latvian since only a minority of Latvian speakers were bilingual in German.

Fifth, the degree of structural similarity of languages at the time of their arrival into an area may also be different and this may represent another obstacle for convergence despite intensive contacts. For example, German has a typologically rare V2 word order which likely emerged already in Proto-West-Germanic or even earlier, i.e., prior to the arrival of German in the CB area. By contrast, Finnic, East Baltic and East Slavic languages must have all originally been SOV, since Proto-Uralic and Proto-Indo-European were both SOV (Janhunen 1982 for Proto-Uralic and Watkins 1963; Dressler 1971; Lehmann 1974 for Proto-Indo-European). Thus, convergence in word order between East Baltic and Finnic languages required much less

[2] The earlier population of the Eastern coastal regions of the Baltic Sea was primarily Finnic and probably West Baltic (but see also Kalio 2015; Lang 2018).

restructuring than convergence of these languages with German. We, therefore, *a priori* expect German to perform differently in our study than Baltic and Finnic languages.

Given that similarity is a subjective and a relative measure, scholars rarely achieve a general agreement on which linguistic traits are areal and which are not (Campbell 2006: 2; "the feature problem" in van Gijn & Wahlström 2023: 179-180). Eventually, it is up to the researcher to arbitrarily decide between the two options: (i) the differences between similar traits in the languages of an area are negligible and, therefore, these traits may be claimed to bear areal effects, and (ii) these traits are rather too different from each other for such a claim. For example, while for Jakobson (1931[1971]: 137), polytonicity is one of the defining traits of the CB area, Wälchli & Koptjevskaja-Tamm (2001: 640-646) are much more cautious. They discuss in detail the emergence of polytonicity in Baltic, Livonian (Finnic) and Scandinavian languages of the CB area and come to the conclusion that these systems are quite distinct in terms of time depth and pathways of their emergence in the three branches and cautiously conclude that it is "(U)nclear whether the three phenomena are related to each other" (Wälchli & Koptjevskaja-Tamm 2001: 729). Jakobson (1931[1971]) was also aware of the same facts but his subjective threshold for claiming contact effects was apparently lower than that of Wälchli & Koptjevskaja-Tamm (2001). Our approach allows avoiding subjectivity.

3. The distance-based approach

To avoid many of these methodological problems, we propose the *distance-based approach.* This approach implements a somewhat different concept of areal convergence, which is not based on similarity among languages, but rather on positive diachronic dynamics *towards* convergence. In our view, convergence, being a complex, never-ending, dynamic process, does not need to result in a high degree of similarity of a phenomenon across the languages of an area. Minimally, a convergent development is just a change (or retention) *towards*, but not necessarily *into*, similarity with the other languages of the area. This concept of convergence better integrates the historical mechanisms by which convergence evolves: different arrival and departure times of languages in and out of an area, varying structural (dis)similarity prior to contact across subsets of languages in an area and thus the number of changes needed to achieve similarity, the total duration of contacts in an area, the specific contact configuration in an area, etc., see §2.

We implement this understanding of convergence into our approach and shift the focus away from providing evidence for similarity of a phenomenon among the languages of an area – a methodological requirement that a priori leaves out many convergent processes which, however, were not yet sufficient to reach a high degree of similarity. Accordingly, our method aims at capturing diachronic dynamics of languages and crucially relies on the question whether or not a language may be said to have developed *towards* the other languages of the area or not. The method is thus not based on similarity judgements.

We operationalize our concept of areal convergence as follows. We suggest that areal convergence is found if it can reasonably be shown that languages undergo (non)-changes away from their close relatives outside the area *towards* the languages of the area (drawing on Di Garbo & Napoleão de Souza 2023). In this respect, the distance-based approach follows more recent approaches such as Ranacher et al. (2021), Di Garbo & Napoleão de Souza (2023) or Sinnemäki et al. (2024), the major differences being that these studies rely on categorical features and global sampling.

We primarily draw on Di Garbo & Napoleão de Souza (2023). Their goal is to approach language contact from a typological perspective in order to explore and produce generalizations about worldwide contact scenarios. For this purpose, they developed a method that we draw on

here. The method allows them to estimate the probability of contact effects on the Focus language by the Neighbor language without going into diachronic research and/or research into the specific contact situation. This method crucially relies on the notion of Benchmark language as a baseline. A Benchmark language is genealogically related to the Focus language but has no contact with either the Focus or the Neighbor language and serves as a *tertium comparationis*. Every potential aspect of similarity of a specific grammatical category between the Focus and the Neighbor language is compared to the Benchmark language. If the Benchmark language also exhibits a similar trait, a contact effect is not supported. However, if the Benchmark language deviates from the Focus language but the Focus language correlates with the Neighbor language then a contact effect from the Neighbor language on the Focus language can be safely assumed. In this way contact-induced similarities between languages are identified. This approach requires, accordingly, a special way of sampling, making sure that the dataset would consist of language triples. This, in turn, brings about limitations, some of which are similar to our approach, e.g., isolates or languages with only distant relationships are problematic because a reasonably justified Benchmark language is not available.

While the focus of Di Garbo & Napoleão de Souza (2023) is to put forward a better way to control for areal biases in typological research as well as typologizing over contact situations, our focus is reverse. We zoom in and seek to explore and better understand linguistic areas, their internal composition, genealogical effects and contact effects on specific linguistic phenomena. We also adopt a more flexible definition of linguistic areas in (1), which allows for linguistic areas consisting (solely) of closely related languages and even (dia)lects of the same language, provided geographical contiguity.

(1) Definition of a sprachbund / linguistic area
A sprachbund represents an idiosyncratic clustering of linguistic traits in a geographical area (containing more than one lect) as opposed to the wider geographical background.

While fine-grained categorical data in principle can be used with the distance-based approach as well, in this paper, our data is token-based extracted from corpus.

More specifically, our approach involves three steps.

First, we identify the set of languages (Focus languages) that have been claimed to belong to a linguistic area, the CB area in this study, see the list of the CB languages above in the Introduction. We take a subset of these languages, for which the relevant corpus data is available (see the list below in Table 1). For each language suggested to be part of the area, we establish its Benchmark language. A Benchmark language is a language that is close to the Focus language both genealogically and geographically, but which does not belong to the area. For example, Dutch is outside of the CB area, but it is genealogically and geographically close to German – a language inside the area. Similarly, Ukrainian (non-CB) is a Benchmark language for Belarusian (CB) and Russian (CB), see Table 1 in section 5. Only Latvian and Lithuanian do not have Benchmark languages, since these languages are the only living Baltic languages. This is a natural limitation to the distance-based approach. Languages within the area must have closely related languages outside of the area. This is not the case with small (sub)families and isolates. This constrains the applicability of the approach. Note, however, that this is just a limitation of the approach and not a constraint on the type of language that may be part of an area.

Second, we test whether the Focus language (e.g., Belarusian) is closer to the languages of the area than its Benchmark language (Ukrainian) with respect to the phenomenon at issue, i.e., word order in this study. If so, then Belarusian is likely to have undergone some areal convergence with respect to word order. We repeat the procedure subsequently for all

languages of the area which have Benchmark languages. We thus identify the set of convergent languages with respect to the specific linguistic phenomenon.

Under the distance-based approach, it is entirely irrelevant whether the areal impact was conservative, i.e. exercising pressure for no change, or innovative, i.e. exercising pressure for change. That is, it is irrelevant whether, say, Belarusian has preserved from Proto-East-Slavic more similarities with the languages of the area than Ukrainian or whether Belarusian developed some innovations towards the patterns of the area which Ukrainian did not. What matters only is that, with the respect to the trait, the difference between Ukrainian and the CB languages is larger than the difference between Belarusian and the CB languages.

Finally, once the set of convergent languages has been established, we may explore the internal composition of the area, based on pairwise (dis)similarities between the languages of the area, technically implemented as pairwise distances.

We summarize these steps in (2), indicating sections of the paper where we discuss the respective steps of our study:

(2) The distance-based approach
Step 1: "Setting up the data" (§5)
Identify the linguistic phenomenon; identify the set of languages (Focus languages) to be tested for areal convergence as well as the set of the Benchmark languages outside of the area.
Step 2: "Identifying convergent languages" (§7)
Explore whether there is a distance between the Focus and its Benchmark such that the Focus is closer to the other languages of the area than its Benchmark.
Step 3: "Exploring internal relations within the convergent languages" (§7)
Explore and explain the degree of similarities across the convergent languages.

4. Word order

In this paper, we focus on the order of words in sentences in the running text. It has been repeatedly emphasized that typological word order patterns such as SVO vary greatly across languages as to their corpus frequencies (Mithun 1987; Dryer 1989; Downing 1995: 19; recently Levshina et al. 2022). Languages vary in the exact conditions of the occurrence of their basic word orders and thus in the frequencies with which these word orders are found in corpora. Different factors may affect word order in a language such as the lexical (animacy, part-of-speech) and discourse (giveness, definiteness) factors (Dryer 1997: 73), information-structural profiles the particular word orders may have (Mithun 1987; Dryer 1989), interactional factors affecting word order such as turn-taking (Downing 1995; Du Bois 2014; "intersubjective coordination" in Verhagen 2005; Tanaka 2005; Selting & Couper-Kuhlen 2000: 86-89), specific effects of more efficient sentence processing (e.g. Seržant et al., forthc., on Russian) and possibly other factors. The combinations and the impact of specific factors and their effect size are obviously language-specific and thus idiosyncratic. Correlations in idiosyncratic traits is a methodological requirement to argue for effects of language contact against the null hypothesis of genealogical effects and/or spread of universally preferred and common traits, see (1) above (Seržant 2015: 330-331; Seržant, aop). Specific corpus frequencies – in contrast to biases (e.g. for OV vs. VO) – can only be language-specific and can neither be universally preferred nor inherited over generations.

Specifically, we explore the frequency and the degree of match in the order of words across sentences on the basis of parallel translations of the same text (Bible) for 10 Circum-Baltic languages. We approach the comparison of the word orders in the 10 languages

somewhat agnostically by comparing the sequences of words in every sentence. We do not directly explore the distribution of typological primitives such as S, V and O as is usually done in large-scale typological works (like Dryer 1989, 1992, 2011), including corpus-based approaches. The typological primitives S, V and O are also unlikely to be sufficient for identifying the points of variation and correlations across our languages as they gloss over many different types of syntactic structures such as complex predications, various non-argumental and oblique object NPs, discourse particles, etc. Our approach thus indirectly captures fine-grained linear differences, which, however, remain to be identified and described in more detail in future work.

5. Data

The study is based on the Parallel Bible Corpus comprising ca. 900 translations into 830 language varieties (Mayer & Cysouw 2014; Plungian 2023; see a collection of papers based on this corpus in Khomchenkova et al., eds., 2023). From this corpus we extracted 16 languages, of which 10 belong to the Circum-Baltic area (Focus languages) and another 6 are CB-external languages which are close to these 10 Focus languages genealogically and geographically and will be used as Benchmark languages.[3] Table 1 summarizes our sample.

Table 1: Languages of our sample

Language	Branch	Family	Part of the area	Benchmark language	Bible translation, metadata[4]
Belarusian	East Slavic	Indo-European	yes	Ukrainian	"Belarusian New Testament and Proverbs." Translated by A. Bokun. 2023.

[3] We relied on previous research that determines which languages are part of the CB area and which are not. However, theoretically our method may be used to provide evidence in favor or against including a language into an area depending on its similarity to the other languages of the area and crucially on its distance to its close relatives that are unequivocally outside of the area, if the evidence from different phenomena would accumulate towards the area. Furthermore, we did not include different dialects and diachronic layers of the same languages for convenience. Thus, Low German is not included into the sample even though this was an important language in the beginning of the Hansa in the region (13-14th c.). This language can easily be added in subsequent research. Low German quite soon ceased to be the main language of Hansa as more and more traders from High German areas became active in Hansa. German trade documents were prevailingly composed in High German since then.

[4] The translations we used are available in the following documents of the corpus: File: pol-x-bible-nowagdansk.txt, Lines: 7958, Tokens: 179804; File: rus-x-bible-modern2011.txt, Lines: 7958, Tokens: 196672; File: lit-x-bible-ecumenical.txt, Lines: 31157, Tokens: 677209; File: bel-x-bible-bokun.txt, Lines: 7958, Tokens: 179454; File: fin-x-bible-1992.txt, Lines: 31170, Tokens: 678920; File: ces-x-bible-bible21.txt, Lines: 31163, Tokens: 704965; File: nob-x-bible-2011.txt, Lines: 7491, Tokens: 194244; File: ukr-x-bible-2009.txt, Lines: 31173, Tokens: 762472; File: eng-x-bible-common.txt, Lines: 7942, Tokens: 210762; File: swe-x-bible-2000.txt, Lines: 35161, Tokens: 878160; File: deu-x-bible-meister.txt, Lines: 7957, Tokens: 209006; File: est-x-bible-1997.txt, Lines: 31173, Tokens: 724525; File: rmy-x-bible-vlax.txt, Lines: 7958, Tokens: 224382; File: nld-x-bible-2007.txt, Lines: 7920, Tokens: 228476; File: lav-x-bible-1997.txt, Lines: 7956, Tokens: 178259; File: pol-x-bible-covenant.txt, Lines: 7956, Tokens: 180128; File: rml-x-bible.txt, Lines: 879, Tokens: 18670; File: hun-x-bible-2012.txt, Lines: 7943, Tokens: 210886.

Czech	West Slavic	Indo-European	no	-	“Czech Bible, 21st century translation.” Biblion (First edition) 2009.
Dutch	West Germanic	Indo-European	no	-	“The Bible in Dutch.” Biblica, Inc. 2007.
Estonian	Finnic	Uralic	yes	Hungarian	“The Bible in Estonian.” Estonian Bible Society. Eesti Piibliselts 1997.
Finnish	Finnic	Uralic	yes	Hungarian	“The Bible in Finnish.” Version of 1992.
German	West Germanic	Indo-European	yes	Dutch	“The New Testament in German.” Abraham Meister Version. 1989.
Hungarian	Ugric	Uralic	no	-	“The New Testament in Hungarian. Simple translation.” World Bible Translation Center. 2012.
Latvian	Baltic	Indo-European	yes	absent	“Revised Latvian Bible”. Revised translation from 1965. Latvian Bible Society. 1997.
Lithuanian	Baltic	Indo-European	yes	absent	“The Bible in Lithuanian Bible, Ecumenical edition.” Bible Society of Lithuania. 1999.
Norwegian	North Germanic	Indo-European	no	Swedish	“The Bible in Norwegian (bokmål).” The Norwegian Bible Society. 2011.
Polish	West Slavic	Indo-European	yes	Czech	“The New Covenant Translation of the Bible in Polish.” Evangelical Bible Institute. 2011.
Russian	East Slavic	Indo-European	yes	Ukrainian	“The New Testament - A modern Translation in Russian.” Corporation World Bible Translation Center. 2011.
Swedish	North Germanic	Indo-European	yes	Norwegian	“The Bible in Swedish.” Swedish Bible Society. 2000.
Ukrainian	East Slavic	Indo-European	no	-	“The Bible in Ukrainian.” 2009.

Baltic Romani	Indo-Aryan	Indo-European	yes	Vlax Romani	"St John's Gospel in Lettish Romani." 1933. British and Foreign Bible Society. 1933, 2016.[5]
Vlax Romani	Indo-Aryan	Indo-European	no	Baltic Romani	"New Testament in Romani." 1984 - Ruth Modrow. Ramosardya pe rhertia pala International Gypsy Publications Inc, Seattle USA.

Unfortunately, Baltic Romani only includes some parts of the entire Bible text (approx. 10%), which means that the comparison of word order for Romani vs. all other CB languages relies only on this part. We think, nevertheless, that given the amount of the text (more than 24.000 verses), this did not affect our analysis.

With the other languages we examined, the entire texts available. We did not perform any adjustments of the amounts of texts used across the language and strived for the maximum text lengths available in the Parallel Bible Corpus. Larger text amounts capture more variation in each of the languages, making sure that our results are more robust against potential text-internal variation.

When it comes to the Bible translations chosen for our study, we purposely selected the newest translations, which are presumably less influenced by the translational tradition and closer to the contemporary language. This was important in order to minimize translational effects of one language on the other. Modern translations aim at better comprehensibility of the Bible text and care much less about preserving older styles which themselves are often biased by the original text (i.e. by languages such as Latin, Greek, Biblical Hebrew, Church Slavonic, etc.).

6. Computation

In order to compare word order across all sentences in the 16 languages, first, the words in each language pair were automatically aligned. This was achieved by employing the eflomal application, which is based on the earlier efmaral tool (Östling & Tiedemann 2016). The eflomal tool automatically aligned words[6] for all matching sentences across the 16 languages. Specifically, we utilized Model 3 of the eflomal tool that is the successor to earlier alignment models and has demonstrated superior performance in preliminary evaluations for several language pairs compared to other models. The tool provides every aligned word pair with the word positions in the sentences. An example of the outcome of eflomal for a sentence pair in English and German (taken from the Bible corpus) is given in Figure 1.

[5] This is one of the earliest Bible translations into Romani. It was translated by a native speaker Janis Lejmanis who was a Latvian Rom and a member of the Orthodox Church. His translation was checked both by an educated Rom from Latvia and by the Scottish scholar Sir Donald MacAlister (1854–1934) (van den Heuvel 2020: 461).

[6] eflomal aligns all tokens in the sentences, including punctuation marks in addition to words. However, for the sake of simplicity, we will refer to the "word" as the minimal unit in the sentence. Note that punctuation or orthography does not influence the results. In the languages under analysis, in order for the punctuation signs to mismatch there needs to be mismatch in word order. We additionally checked whether the distances calculated based on words only and not taking into account punctuation marks differ from those we use in the paper. We found only no differences and a perfect correlation of 1 between the two matrices, using Mantel test from the package vegan (Oksanen et al. 2025).

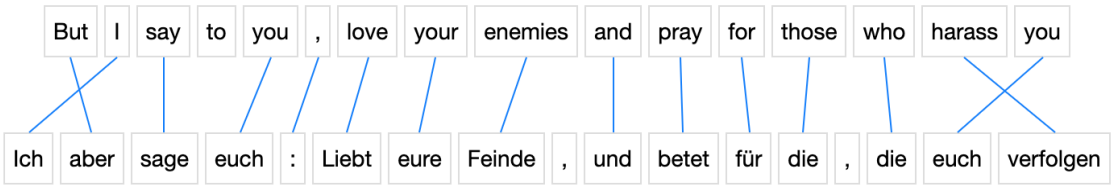


Figure 1: An example for the world alignment in the German and English Bible translations

In this sentence, there is a word pair *I-Ich,* with the positions 1-0, meaning that *I* is the second word in the English sentence and *Ich* is the first word in the German sentence (with numbering starting from 0). The crossing lines in Figure 1 indicate that two words in the English sentence (i.e., *But* and *harass*) are not in the same linear order as their correspondences in German. However, all other words maintain the same linear order. In this case, the metric for this sentence for the pair German-English would be calculated by dividing 13 (the number of words in the same linear order) by 15 (the total number of words in the English sentence that are aligned with words in the German sentence). Thus, the score for this sentence pair is 0.87. Once we computed the scores for all aligned sentences and language pairs, we calculated the average scores of word-order matches for each language pair. These average scores were then used to represent mutual word-order distances between each pair of the languages.

In case of periphrastic constructions in which, for example, a preposition plus a lexical noun, say *mit* 'with' plus NP in German correspond to one word in the other language, say, the noun in the instrumental case in Belarusian, the algorithm aligns only the lexical correspondence and ignores the function word. This is found with *to* in the English sentence in Figure 1 which is rendered in the dative-marked pronoun in German and has, therefore, no alignment pair. In this way the distinction between periphrastic vs. synthetic forms does not influence the alignment and, thus, the differences between more synthetic languages like Estonian vs. more analytic languages like German cannot affect our results.

7. Results[7]

As the first step in our analysis, we estimated whether each Focus language of the CB area is closer to the other CB languages than its Benchmark language, with respect to the order of words in sentences. Then, we tested whether these differences, technically distances, between each CB language and its Benchmark language are statistically significant. This allowed us to estimate whether the languages of the Circum-Baltic area form a cluster as opposed to the surrounding languages, cf. (1) above.

As an example, consider the pair of Polish, which belongs to the CB area, and Czech, which is its Benchmark language. Czech is close to Polish both genealogically (both belong to West Slavic) and geographically, but Czech does not belong to the CB area. Table 2 shows the distances between each of these two languages and the other languages of the area.

Table 2. Distances of Polish and its Benchmark language Czech to the languages of the Circum-Baltic area (except Polish)

	Baltic	Belaru	Estoni	Finnis	Germa	Latvia	Lithua	Russia	Swedis

7 The data and the code for the analysis and visualizations discussed in this section are available at Aktaş et al. 2025.

	Romani	sian	an	h	n	n	nian	n	h
Polish	0.08	0.11	0.14	0.15	0.13	0.13	0.12	0.12	0.14
Czech	0.15	0.15	0.17	0.18	0.17	0.13	0.16	0.14	0.17

Except for Latvian, Polish is consistently closer to the languages of the CB area than its Benchmark language Czech. To check the statistical significance of these differences, for each pair of the CB languages and their Benchmarks, we used Wilcoxon signed-rank test. Table 3 shows the mean distance of each language of the area to the remaining languages of the area and the mean distance of its Benchmark language to the same set of languages. It also shows the *p*-values and the effect size[8] obtained as a result of the tests for each pair under comparison (the sample size, i.e., the number of paired distances under comparison, always equaled eight). Two of the languages outside of the area, Hungarian and Ukrainian, are used as Benchmark languages twice, and have two mean values, depending on which language of the area they are compared with. Baltic languages do not have a Benchmark language outside of the area, but the mean distance to the remaining languages of the area is also given for them.

Table 3. Mean distances of the languages of the area and their Benchmark languages (if available) to the other languages of the area and the results of Wilcoxon signed-rank test (p-values and the effect size) for each pair of languages of the Circum-Baltic area and their Benchmark languages

Language of the Circum-Baltic area	Mean distance to the other CB-languages	Benchmark language	Mean distance to the other CB-languages	p-value	Effect size
Baltic Romani	0.08	Vlax Romani	0.17	0.009	0.89
Belarusian	0.12	Ukrainian	0.16	0.009	0.89
Estonian	0.12	Hungarian	0.21	0.009	0.89
Finnish	0.13	Hungarian	0.21	0.009	0.89
German	0.14	Dutch	0.26	0.004	0.89
Polish	0.12	Czech	0.16	0.01	0.87
Russian	0.12	Ukrainian	0.16	0.01	0.85
Swedish	0.13	Norwegian	0.14	0.3	
Latvian	0.11	*	*	*	*
Lithuanian	0.11	*	*	*	*

[8] The effect sizes for the Wilcoxon signed-rank tests were calculated using the finction *wilcox_effsize* of the R package *rstatix* (Kassambara 2023).

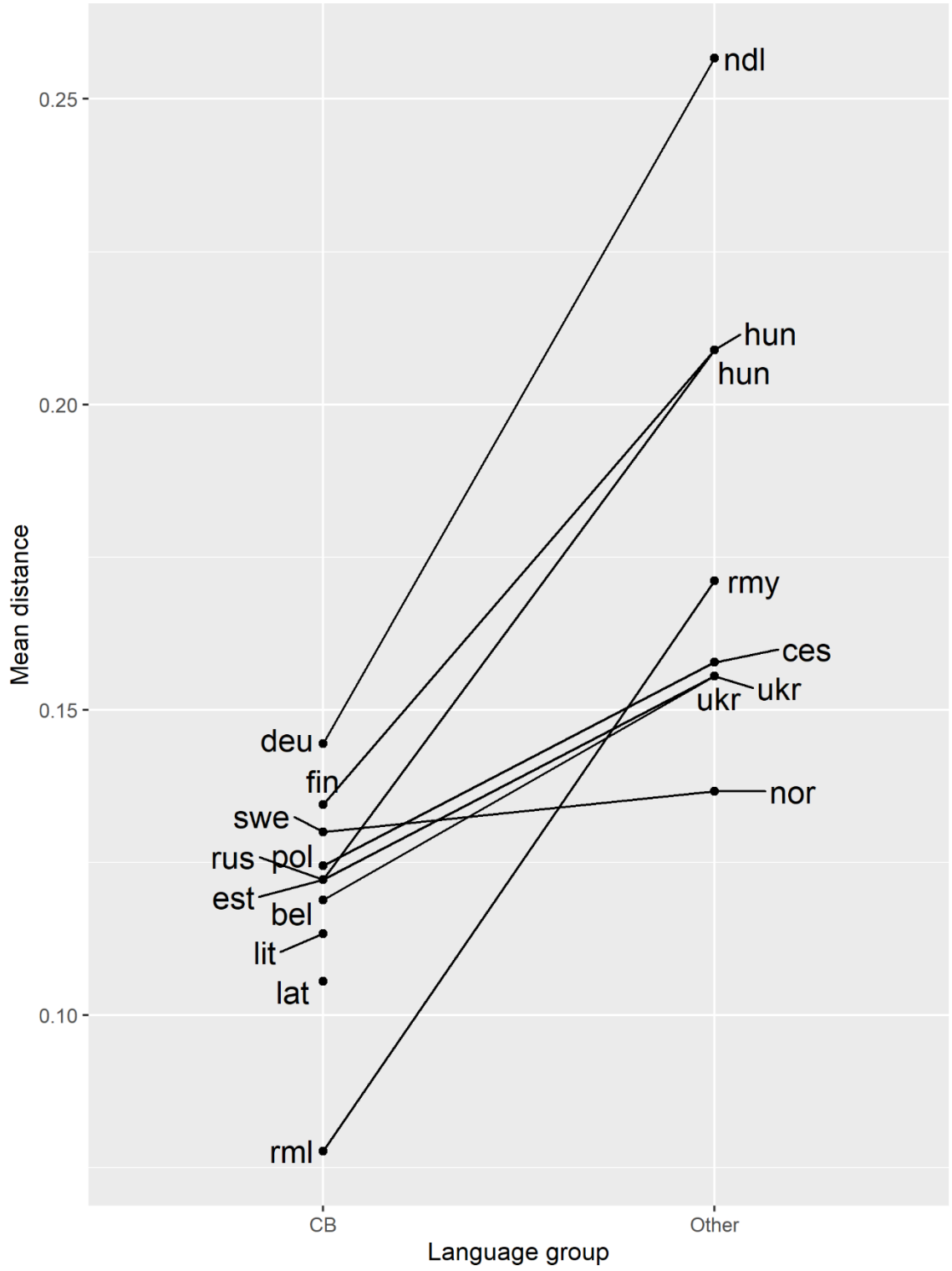
Table 3 shows that the languages of the CB-area have smaller distances to the other CB-languages under comparison as compared to their Benchmark languages. The differences for all pairs are statistically significant with a large effect size, with the exception of Swedish and Norwegian, where no significant difference was found.

The only likely explanation of these non-accidental and substantial differences between the Focus and Benchmark languages is that the Focus languages underwent convergence with the other languages of the area (Step 2 in (2) above). From this it follows that the CB languages have undergone historical changes (or non-changes) which resulted in them being closer to each other in contrast to their Benchmark languages. At the same time, this approach allows the CB languages to be distinct from each other; what matters only is that the CB languages show a statistically significant difference towards the other languages of the area.

The mean distances given in Table 3 are visualized in Figure 2, which represents the difference between the CB-languages and the Benchmark languages graphically.

Figure 2. Mean distances between the languages of the sample to the CB languages[9]

[9] In Figure 2, as well as in Table 4 and Figures 3 and 4 below, we use the ISO 639-3 abbreviations for the languages: *bel* – Belarusian, *ces* – Czech, *deu* – German, *est* – Estonian, *fin* – Finnish, *hun* – Hungarian, *lat* – Latvian, *lit* – Lithuanian, *nld* – Dutch, *nor* – Norwegian, *pol* – Polish, *rml* – Baltic Romani, *rmy* – Vlax Romani, *rus* – Russian, *swe* – Swedish, *ukr* – Ukrainian.



The steepness of the lines in Figure 2 can also be interpreted in terms of the amount of change towards convergence of each of the languages to the other members of the CB area as well as the effect of the genealogical and geographical distance between Focus and Benchmark languages. It is not easy to disentangle these two effects. Thus, Hungarian – as opposed to more closely related languages such as Erzya or Moksha (not according to Glottolog) – is more distant from both Finnic CB languages Finnish and Estonian, both geographically and genealogically. These languages also have the largest distance between Focus and Benchmark.

However, Dutch and German are closely related both genealogically and geographically but nevertheless show a large distance between Focus and Benchmark, which indicates that German is highly convergent. Even though being highly convergent, it is, nevertheless, as predicted above in Section 2, a language that is still the most distant from the other languages in the area (average distance is 0.14). This is an effect that might be explained by quite a special word order of German prior to contact.

Swedish is the least convergent language as it deviates from its Benchmark language Norwegian towards the languages of the CB area only very slightly. This evidence supports previous claims about the CB area that the Eastern part thereof is subject to more intensive contact effects than the entire area, which also includes Scandinavia.

Given that Baltic Romani shows the least distance to all the other CB languages (0.08) and that it has one of the highest distance to its Benchmark, we may conclude that Baltic Romani quite intensively developed towards the CB area and it is one of the most convergent languages in the area. One of the reasons why Baltic Romani is highly convergent might be sought in the fact that Romani dialects are not standardized languages and are subject to language ideology and prescriptivism to a much lesser extent than the other languages of our sample, which are official languages in the respective states. Thus, Romani dialects seem to be generally more flexible in adapting word order traits of their contact languages. Matras (2002: 167-169) lists a number of innovations in word order Romani dialects adopted from their neighbors. For example, Matras (2002: 168) notes that, under Slavic influence, some Romani dialects acquired a new "tendency to place the object, and especially the pronominal object, before the verb."[10] By contrast, Sinti varieties have adopted the German word order to a different extent; Romani dialects in Azerbaijan and Turkey tend towards verb-final order, as in Western-Oghuz Turkic (Matras 2002: 168).

Now we turn to Step 3 of our approach, as described in (2), and explore the internal composition of the CB languages with respect to the order of words. Here, we no longer focus on the dissimilarity of the Focus languages with their Benchmark languages but rather highlight the similarities among the CB languages since they have been shown to be convergent to the area with respect to word order in Step 2.

First, we explore the structure of the area by looking at the mean distances of the CB languages to each other. The overall picture of word-order distances in the CB and non-CB languages under scrutiny is found in Table 4. This table shows one half of the distance matrix only, as it is symmetrical about the diagonal. The darkness of the shading corresponds to the distance value. The Circum-Baltic languages are grouped in the left part of the table.

Table 4. Distances between the languages

[10] Pronominal objects prefer OV in Russian, for example, despite the fact that this language is generally VO (Seržant et al., forthc.).

		CB										Other					
	bel	deu	est	fin	lav	lit	pol	rml	rus	swe	ces	hun	nld	nob	rmy	ukr
bel	0	0,14	0,15	0,15	0,14	0,11	0,11	0,04	0,09	0,14	0,15	0,22	0,30	0,15	0,17	0,11
deu		0	0,14	0,20	0,12	0,15	0,13	0,09	0,17	0,16	0,17	0,25	0,26	0,17	0,19	0,16
est			0	0,11	0,08	0,11	0,14	0,09	0,14	0,14	0,17	0,20	0,29	0,13	0,20	0,18
fin				0	0,09	0,14	0,15	0,09	0,13	0,15	0,18	0,20	0,31	0,13	0,21	0,20
lav					0	0,08	0,13	0,08	0,13	0,10	0,13	0,20	0,18	0,12	0,11	0,15
lit						0	0,12	0,07	0,10	0,14	0,16	0,21	0,30	0,14	0,20	0,18
pol							0	0,08	0,12	0,14	0,11	0,23	0,21	0,14	0,13	0,14
rml								0	0,09	0,07	0,15	0,12	0,16	0,10	0,08	0,11
rus									0	0,13	0,14	0,22	0,29	0,15	0,17	0,13
swe										0	0,17	0,23	0,27	0,07	0,16	0,17
ces											0	0,24	0,28	0,16	0,22	0,19
hun												0	0,29	0,22	0,22	0,23
nld													0	0,25	0,26	0,32
nob														0	0,16	0,16
rmy															0	0,20
ukr																0

Here, we also see that Baltic Romani is generally very close to the other languages of the CB area. Furthermore, Latvian (Indo-European, Baltic) is as close to its close relative Lithuanian as to Estonian (0.08) (Finnich, Uralic). Latvian is very close not only to its next relative Lithuanian (0.08) but also to the genealogically unrelated Estonian (0.08). This specific contact configuration of Latvian is motivated historically. This language was closely affiliated with Estonian for political reasons both during the time of Livonia (founded by the Teutonic Order by the end of the 12th c.) as well as later under the Swedish reign (up until 1721). In turn, Lithuanian was part of the Polish-Lithuanian Commonwealth (up until 1795) together with Belarusian and Polish and it, indeed, shows lower distances to these languages. German is on average more distant to the languages of CB area than the other languages, which is also expected given that this language was only dialectally and as a superlect present in the CB area. Languages that are the closest to German (within CB) are: Latvian (0.12), Estonian (0.14), Polish (0.13) and Baltic Romani (0.09).[11] These results lend support to our approach as we find smaller distances between those languages for which we independently know about their more intensive contacts.

To visualize these distances, we used a Multidimensional Scaling algorithm, as implemented in the package *smacof* (Mair et al. 2022) in R (R Core Team 2024). This dimensionality-reduction method is used to represent distances between objects in a 2- or 3-dimensional space, aiming at a minimal distortion of the distances. The degree of distortion is expressed in a value called stress, with the acceptable values of stress being lower than 0.20 (Levshina 2015: 341). The 2-dimensional visualization of our distance matrix in Figure 3 has the stress value 0.24, which is above the acceptable threshold. However, the 3-dimensional visualization in Figure 4, which has an acceptable stress value of 0.17, shows basically the same picture, with the languages of the Circum-Baltic area being closer to each other, and the other languages at the periphery of the graph.

Figure 3: The 2-dimensional MDS-plot visualizing word-order distances between the languages

[11] Interestingly, we do not see much of the effect of written Estonian being largely also V2 like German (Vihman & Walkden 2023) in our data. This might have many different explanations. For example, the verb-second position is natural in any SVO language and might be the most frequent one in any of the languages of the area which is why Estonian does not play out to be more closely associated with German than other languages.

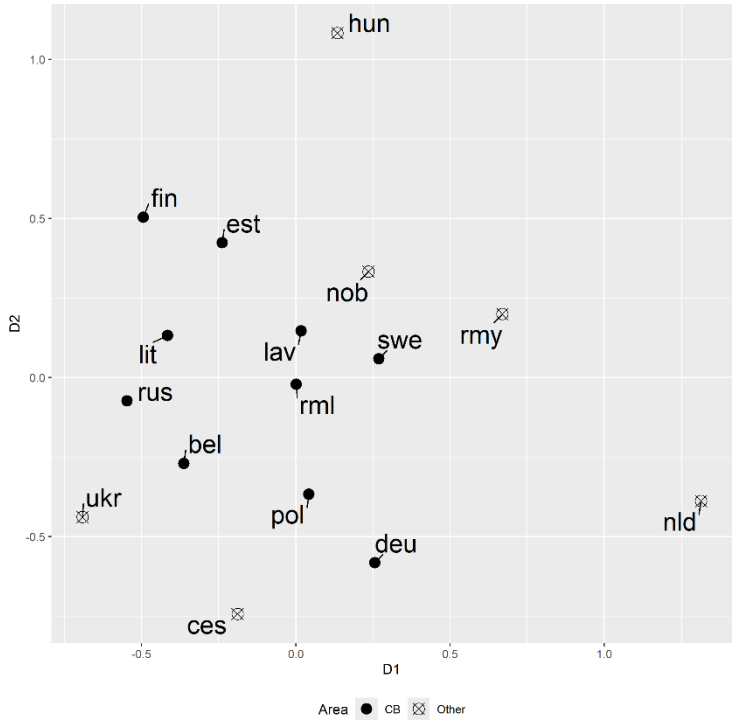
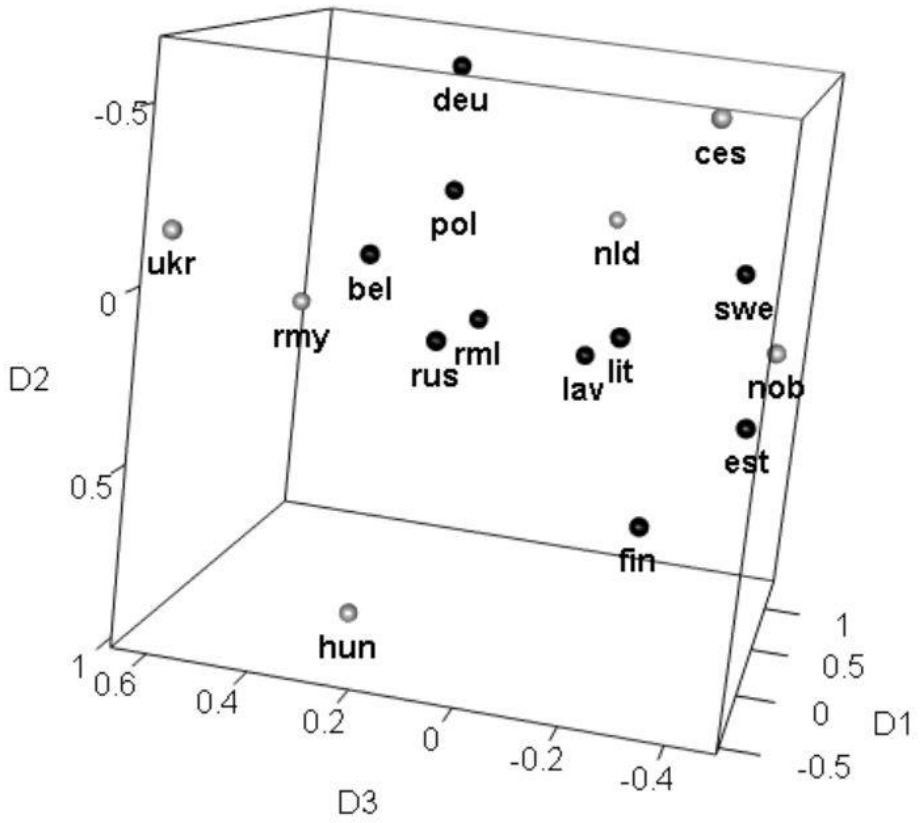


Figure 4: The 3-dimensional MDS-plot visualizing the word-order distances between the languages



What we see in Table 4, Figures 3 and 4 is that the languages of the Circum-Baltic area (marked with black dots) generally cluster closer to each other than to their Benchmark languages (crossed circles in Figure 3, grey dots in Figure 4). MDS thus renders the areal effect of convergent developments (cf. Matras & Sakel 2007) in the CB area.

In addition to the areal effects, the genealogical signal is very strong, since many closely related languages are found next to each other in Figure 4: the three East Slavic languages (Ukrainian, Belarusian, Russian), and, not far from them, Polish (which is a West Slavic language) are placed next to each other.[12] Likewise, Scandinavian languages (Norwegian and Swedish), and the two Finnic languages (Finnish and Estonian) pattern close to each other. The two Baltic languages Latvian and Lithuanian are very close to each other as well.[13] However, the row numbers in Table 4 above unequivocally show this: the distance between Lithuanian and Latvian is the smallest (0.08) for Lithuanian and it is also the smallest for Latvian, which, however, in addition, has the same minimal distance with Baltic Romani and Estonian.

To support our observations on the areal convergence between the CB languages, we tested for the correlation between the distances based on word order and the membership in the CB area, controlling for the genealogical affiliation. To do this, we used partial Mantel test, which checks for the presence of correlation between two matrices (in our case, word-order

[12] Note that Polish - in contrast to the other West Slavic languages such as Czech - patterns with East Slavic also in other traits, for example, in argument marking (see Seržant et al. 2022) or aspect (Dickey 2000).

[13] Recall that MDS is a 3-dimensional representation of an originally multidimensional space so some information might be lost in the visualization. This is the case in two-dimensional Figure 3 as opposed to three-dimensional Figure 4 where both languages are close to each other on a dimension that disappears from Figure 3.

distances and membership in the area), controlling for the factor given in the third matrix (genealogical "distance"). We encoded the membership in the area as 1 for the pairs of CB languages and 2 for all other pairs, and the genealogical relations as 1 for pairs belonging to the same subfamily (such as Slavic or Finnic), 2 for the pairs of languages from different subfamilies within the same family, and 3 for the languages from different families. The test showed a significant correlation between the word-order distances and the membership in the area (Mantel statistic r = 0.57, p = 0.002). We also tested for the significance of the correlation between the distance matrix and the matrix encoding genealogical relations, controlling for the areal factor, i.e., for the membership in the Circum-Baltic area. In this case, the partial Mantel test shows a weaker correlation between the matrices, which is only marginally significant (Mantel statistic r = 0.27, p = 0.06). Thus, within our language sample, we see strong evidence for the areal convergence of Circum-Baltic languages, accompanied by a more moderate genealogical effect.

However, as we argued in Section 2 above, it is doubtful that within an area of intensive language contact, the genealogical signal is solely conditioned by the independent preservation of commonly co-inherited traits. We suggest that *genealogical pressure* and *genealogical signal* should not be equated but rather – given also the geographical proximity between the genealogically related languages – are notionally in the subset-superset relation (Seržant, forthc.). It is more likely that genealogical signal results from the combination of two independent diachronic mechanisms: (i) coincidental preservation of a very large amount of commonly inherited traits (*genealogical pressure*) and (ii) strong effects of *language contact* enhanced by the high degree of similarity between the closely related languages (due to the very large amount of commonly inherited traits), leading to a more intensive exchange in innovations (see Epps et al. 2014). These effects may lead either (ii.a) to non-change and preservation or (ii.b) to common innovations (see Seržant 2021 for a case of contact-induced preservation vs. contact-induced innovation). While we do not have a specific example from the domain of word order given the quantitative nature of our study, there are several common innovations in the area which expanded along genealogical nodes, thus boosting the genealogical signal, which, however, are not due to inheritance. For example, in the East Slavic languages, the new perfect construction based on the invariant active past participles in *-vši* is found in both Russian (3) and Belarusian (4) dialects (Trubinskij 1984; Erker 2014; Pozharickaja 2014).

(3) Western Russian dialects
Rebenok prosnu-vši.
child.NOM.SG wake_up-**PRF**
'The child is awaken.' (Pozharickaja 2014: 112)

(4) Belarusian dialects
fs'a úlica bylá zγaré-ųšy
all.NOM.F street.NOM.F AUX.PST.F.SG burn-**PRF**
'The entire street was burned down.' (Erker 2014: 138)

However, this construction has demonstrably appeared much later than the split of Proto-East-Slavic into Belarusian, Ukrainian and Russian and thus does not represent an instance of inheritance from the common proto-language (Proto-East-Slavic) (Trubinskij 1984: 171).

Another example is the expansion of the prepositional phrase *u* 'at' plus the genitive with the function of *free affectee* in all the three East Slavic languages (traditionally labeled *external possessors*, see Seržant 2016). This is not a phenomenon that one would find in the early Old Rusian documents and, therefore, cannot be claimed to be inherited. Instead, this is

an innovation that spreads via language contact among closely related East Slavic languages and dialects. Thus, contact-induced innovations may be channeled by the genealogical tree but these do not represent genealogical pressure, i.e. inheritance (Seržant, forthc.).

To conclude, even though we see a strong genealogical signal in our data, it cannot be directly interpreted by claiming that the genealogical pressure here is stronger than the areal pressure. Overall, the genealogical signal – due to both the genealogical pressure and the channeling of contact effects along genealogical tress – blurs the overall effect of convergence.

8. Conclusions

We have argued that traditional similarity-based approaches to linguistic areas have difficulties with genealogically more homogenous linguistic areas, with convergence of universally preferred traits and with the subjective degree of similarity sufficient to claim an areal trait. Such approaches do not enable adequate descriptions of linguistic areas and mostly describe areas via typological and areal rara, which are often rara in the corpus of a language as well. These shortcomings are primarily rooted in the methodological difficulties in providing strong evidence against the null hypothesis as well as in the history of areal linguistics. Similarity-based approaches introduce a lot of uncertainty and subjectivity into areal linguistics and research on language contact by requiring that a specific trait in different languages of an area be very similar without a clearly defined baseline. Less similar but still convergent traits cannot be taken into account in these approaches.

Our goal is to shift the focus from comparing similarities in the languages of an area to exploring their dynamics. Our concept of convergence, accordingly, is dynamic and because of that it is not dependent on a high degree of similarity of a comparable trait across the languages of the area. Instead, convergence is understood here as historical processes and pressures that push a language towards the other languages in an area. While we cannot access the specific historical (non-)changes that a language in an area underwent, we approach dynamics indirectly. Technically, we estimate the dynamics based on the distances a Focus language and its Benchmark language have to the other languages of the area (following Di Garbo & Napoleão de Souza 2023). We define the Benchmark language as a language that is geographically and genealogically close to the Focus language but is outside of the area.

Our distance-based approach departs from the assumption that an exact match of two similar traits in two languages is never found and, therefore, what only matters is whether or not the languages of the area exhibit a trend towards convergence. An exact threshold for the degree of convergence or let alone for high similarity across the languages of the area are not required here at all. Finally, our approach is independent of whether the languages of the area are closely genealogically related or not.

We exemplified our approach by exploring the order of words in the languages of the CB area. All languages of the CB area enjoy quite lax constraints on word order. Even the most constrained language German allows for a whole set of word order variants within its V2 and other rules. On the distance-based approach, we found that all languages show convergence effects in their word orders.

We have shown that all languages of the CB area indeed show convergence effects (Step 2). In the next step (Step 3), we compared the CB languages among themselves and found that Baltic Romani and both Baltic languages are in the center of the area by exhibiting the least distance to the other languages of the CB area and thus the highest degree of convergence. Both Baltic languages lack Benchmark languages. However, their distance from the other CB languages (0.11) is the second-lowest distance (Figure 2 and Table 3) among the other CB

languages which, in turn, were shown to undergo convergence effects as opposed to their Benchmark languages.

When it comes to the internal composition and convergence degrees, we also found that Baltic Romani is the most adapted language in the CB area given its high distance to its Benchmark. Latvian and Lithuanian are in the middle. By contrast, German, Finnish and Swedish are the least integrated languages despite the fact that German must have underwent quite some (non-)changes towards the area.

While the specific ways to measure the distance and dissimilarity may vary (see, for example, different approaches in the contributions in Borin & Saxena 2013), the approach we exemplified is designed to be more resistant to differences among languages of an area than similarity-based approaches. The distance-based approach is fine-grained enough to handle even closely related languages in an area, especially if it is applied on corpus data. It can also handle varying degrees of similarities across the languages of the area.

When it comes specifically to word order in the CB area, our approach does not allow us to identify which specific constructions and discourse strategies were adapted through contact. We only see the overall effect. Our study is just the first step to holistically analyze similarities in word order across the languages of the CB area. The next step in the future would be identifying specific discourse moves, properties of the input, interactional and other effects, which affect the choice of constructions, to pin down the specific constructions that these languages share, e.g. topicalization constructions, animacy and/or definiteness-driven placement of arguments, turn-taking effects on word order, etc. This will certainly first require explorative manual analysis and preprocessing of the aligned sentences we have produced for this study. Such an approach would allow aggregating over different types of syntactic and discourse variables in our pre-processed corpus data but would also be very time consuming, since most of such variables would have to be tagged manually. It would also require more difficult computation given that our parallel data are neither syntactically nor part-of-speech-wise tagged. Finally, we also remain agnostic as to the exact diachronic mechanisms that led to convergence in word order. Language external reasons such as social and political history of the region would suggest that it is the Baltic languages that have adapted more to the dominant languages such as German or Russian than vice versa. The same applies to Baltic Romani.

Another aspect that we did not discuss in detail is varying degree of genealogical relatedness between the Focus and the Benchmark languages. Benchmark languages will often be subject to convenience sampling due to lack of specific data. Some languages do not have very close relatives at all (like Baltic in our case) while others do, yet other languages do have close relatives but there is no parallel corpus data available for these (like Erzya and Moksha which would have been better Benchmark languages for Finnish and Estonian). This is an issue that Di Garbo & Napoleão de Souza (2023: 581-582) also address and suggest including a measure of relatedness into the computation. For example, one way to do so might be by means of building a model that would incorporate genealogical distance (cf. Jaeger et al. 2011). Additionally, one might balance the effect of sampling and take more than one Benchmark language to rule out language-specific noise.

Languages with no Benchmark languages may also be included in Step 3 in (2) by comparing their distances to the mean of the languages of the area with such distances of the languages that have been shown to belong to the area on the basis of the distance-based approach. We applied this procedure with respect to both Baltic languages and were thus able to include languages with no Benchmarks.

In addition to the “pure” convergence effects for the languages for which genealogical pressure can be excluded due to their distinct genealogies (such as Polish-Lithuanian or Latvian-Estonian convergence), we also observe a strong genealogical signal. However, we

argued that the genealogical signal should not be oversimplified and equated with genealogical pressure (cf. Seržant, forthc.). To the contrary, close genealogical relationship may channel language contact as well.

Finally, although we provided evidence for convergence in the domain of word order in the languages of the CB area, our study has no bearing on the more general questions of whether or not a linguistic area is a phenomenon *sui generis* that is distinct from just a set of binary contact effects between the neighboring languages (see Dedio et al. 2019; Ranacher et al. 2021 on further discussion). In other words, it remains to be an empirical question whether there is something like linguistic areas with statistical peaks that would distort larger macroareal clines or whether any random set of contiguous languages on macroclines would show an areal effect similar to the one we found in word order in the CB area. This is certainly something that can only be meaningfully explored on the basis of a set of mutually independent phenomena and on a large scale such as Western Eurasia.

Abbreviations (in addition to Leipzig Glossing Rules)

AUX – auxiliary
CB – Circum-Baltic
PRF – perfect
V2 – verb-second

Acknowledgements

We cordially thank Robert Östling for helping us with the corpus. We also thank the two anonymous reviewers, the editor as well as Peter Arkadiev for their numerous and very valuable comments. We, furthermore, acknowledge the financial support by the Deutsche Forschungsgemeinschaft (Project-ID 317633480 – SFB 1287 “Limits of variability”).

References

Aikhenvald, Alexandra Y. & R. M. W. Dixon. 2001. Introduction. In: Alexandra Y. Aikhenvald & R. M. W. Dixon (eds), *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*. Oxford: Oxford University Press. 1–26.

Aktaș, Berfin, Maria Ovsjannikova & Ilja A. Seržant. 2025. Data & scripts for the paper 'Distance-Based Approach to the Circum-Baltic Area.' by Ilja A. Seržant, Berfin Aktaș, Masha Ovsjannikova, Manfred Stede. Studies in Language [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15727333

Borin, Lars & Anju Saxena, eds. 2013. *Approaches to measuring linguistic differences.* Berlin: de Gruyter.

Bowern, Claire. 2013. Relatedness as a factor in language contact. *Journal of Language Contact* 6. 411–432.

Bužarovska, Eleni. 2020. The Contact Hypothesis Revised: DOM in the South Slavic Periphery. *Journal of language contact* 13, 57–95.

Campbell, Lyle. 1985. Areal Linguistics and Its Implications for Historical Linguistic Theory. In: Jacek Fisiak (ed.), *Proceedings of the Sixth International Conference of Historical Linguistics*. Amsterdam: John Benjamins. 25–56.

Campbell, Lyle. 2006. Introduction. In: Yaron Matras, April McMahon & Nigel Vincent (eds.), *Linguistic Areas Convergence in Historical and Typological Perspective*. New York: Palgrave Macmillan. 1-31.

Civjan, T.V. 1979. *Sintaksičeskaja struktura balkanskogo jazykovogo sojuza.* Moscow: Nauka.
Dahl, Östen & Maria Koptjevskaja-Tamm. 1992. *Language typology around the Baltic Sea: a problem inventory*. Papers from the Institute of Linguistics. Stockholm: University of Stockholm.
Dedio, Stefan, Peter Ranacher & Paul Widmer. 2019. Evidence for Britain and Ireland as a linguistic area. *Language* 95(3). 498–522.
Di Garbo, Francesca & Ricardo Napoleão de Souza. 2023. A sampling technique for worldwide comparisons of contact scenarios. *Linguistic Typology* 27(3). 553–589.
Dickey, Stephen M. 2000. *Parameters of Slavic aspect. A cognitive approach*. Stanford
Downing, Pamela. 1995. Word order in discourse: By way of introduction. In: Pamela Downing & Michael Noonan (eds), *Word order in discourse*. Amsterdam: Benjamins. 1-28.
Dressler, Wolfgang. 1971. Zur Rekonstruktion der indogermanischen Syntax. Kuhns Zeitschrift 85. 5–22.
Dryer, Matthew S. 1989. Discourse-Governed Word Order and Word Order Typology. *Belgian Journal of Linguistics* 4(1). 69 – 90.
Dryer, Matthew. 1992. The Greenbergian word order correlations. *Language* 68. 81–138.
Dryer, Matthew. S. 1997. On the six-way word order typology. *Studies* in *Language* 21(I). 69-103.
Dryer, Matthew. 2011. The evidence for word order correlations. *Linguistic Typology* 15(2). 335–380.
Dryer, Matthew S. 2013. Order of Subject, Object and Verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.), WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/81, Accessed on 2024-07-10.)
Du Bois, John. A. 2014. Totwards a dialogic syntax. *Cognitive Linguistics* 25(3), 359-410.
Emeneau, Murray B. 1956. India as a Linguistic Area. *Language* 32. 3–16.
Epps, Patience, John Huehnergard and Na'ama Pat-El. 2013. Introduction. Contact Among Genetically Related Languages. *Journal of Language Contact* 6. 209–219.
Erker, Aksana. 2014. Ways of Expressing the Past Tense in Belarusian Mixed Subdialects Spoken in the Baltic-Slavic Contact Zone. In: Seržant, Ilja A. & Björn Wiemer (eds.), *Contemporary Approaches to Dialectology: The area of North, Northwest Russian and Belarusian vernaculars*. Bergen: John Grieg AS. 130-149.
Gijn, Rik, van. 2020. Separating layers of information: the anatomy of contact zones. In: Smith N., Aboh E.O. & Veenstra T. (eds.), *Advances in Contact Linguistics: In honour of Pieter Muysken.* Contact Language Library no. 57. Amsterdam: John Benjamins. 162-178.
Gijn, Rik, van & Max Wahlström 2023. Linguistic areas. In: E. van Gijn, H. Ruch, M. Wahlström & A. Hasse (eds.), *Language contact: bridging the gap between individual interactions and areal pattern.* Berlin: Language Science Press. 179-219.
Gumperz, John J. & Robert Wilson. 1971. Convergence and creolization: a case from the Indo-Aryan/Dravidian border in India. In Hymes, Dell H. (ed.), *Pigdinization and creolization of languages*, 151-167. Cambridge: Cambridge University Press.
Haig, Geoffrey. 2001. Linguistic diffusion in present-day east Anatolia: From top to bottom. In: Alexandra Y. Aikhenvald & Robert M.W. Dixon (eds.), *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*. 195-224. Oxford: Oxford University Press.
Haspelmath, Martin. 2001. The European linguistic area: Standard Average European. In: Haspelmath, Martin & König, Ekkehard & Oesterreicher, Wulf & Raible, Wolfgang

(eds.), *Language typology and language universals.* (Handbücher zur Sprach- und Kommunikationswissenschaft) Berlin: de Gruyter, 1492-1510.
Heeringa Wilbert & John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change* 13(3). 375–400.
Jaeger, Florian, Peter Graff, William Croft and Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15(2). 281–319.
Jakobson, Roman. 1931[1971]. Über die phonologischen Sprachbünde, *Travaux du cercle linguistique de Prague* 4, 234–240. Cited after the reprint in: Jakobson, Roman. 1971. *Selected writings I: Phonological Studies.* The Hague: Mouton. 137–143.
Janhunen, Juha. 1982. On the structure of Proto-Uralic. *Finno-Ugrische Forschungen* 44. 23–42.
Kallio, Petri. 2015. The language contact situation in prehistoric Northeastern Europe. In: Robert Mailhammer, Theo Vennemann and & Birgit Anette Olsen, (eds.), *The Linguistic Roots of Europe*. 77–102. Copenhagen: Museum Tusculanum Press, University of Copenhagen. 77–102.
Khomchenkova, Irina A., M. D. Vojejkova, N. M. Zaika, M. L. Kisilier, G. A. Mol'kov & A. Ju. Urmanchieva (eds.), *Studies in the theory of grammar, issue 9. The parallel corpus as a grammar database and the New Testament as a parallel corpus.* Acta Linguistica Petropolitana. Transactions of the Institute for Linguistic Studies. Vol. 19 part 3.
Koptjevskaja-Tamm, Maria & Bernhard Wälchli. 2001. The Circum-Baltic languages: An areal-typological approach. In: *The Circum-Baltic Languages. Typology and Contact. Vol. 2. Grammar and Typology*, Östen Dahland Maria Koptjevskaja-Tamm (eds.), Amsterdam, Philadelphia: John Benjamins. 615-750.
Lang, Valter. 2016. Early Finnic-Baltic contacts as evidenced by archaeological and linguistic data. *Journal of Estonian and Finno-Ugric Linguistics* 7. 11–38.
Lehmann, Winfred P. 1974. *Proto-Indo-European Syntax.* Austin: The University of Texas Press.
Levshina, Natalia. 2015. *How to do Linguistics with R: Data exploration and statistical analysis*. Amsterdam, Philadelphia: John Benjamins.
Levshina, Natalia, Namboodiripad, Savithry, Allassonnière-Tang, Marc, Kramer, Mathew, Talamo, Luigi, Verkerk, Annemarie, Wilmoth, Sasha, Rodriguez, Gabriela Garrido, Gupton, Timothy Michael, Kidd, Evan, Liu, Zoey, Naccarato, Chiara, Nordlinger, Rachel, Panova, Anastasia and Stoynova, Natalia. Why we need a gradient approach to word order. *Linguistics* 61(4), 825-883.
Maddieson, Ian. 2013. Tone. In: Dryer, Matthew S. & Haspelmath, Martin (eds.), WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/13, Accessed on 2024-07-09.)
Mair, P., Groenen, P. J. F., De Leeuw, J. 2022. More on multidimensional scaling in R: smacof version 2. *Journal of Statistical Software* 102(10), 1-47.
Mathiassen, Terje. 1985. A discussion of the notion 'Sprachbund' and its application in the case of the languages in the eastern Baltic area, *International Journal of Slavic Philology* 21/ 22, 273–281.
Matras, Yaron. 2002. *Romani: a linguistic introduction*. Cambridge: Cambridge University Press.
Matras, Yaron. 2007. The borrowability of structural categories. In Yaron Matras and Jeanette Sakel (eds.), *Grammatical Borrowing in Cross-linguistic Perspective*. Amsterdam: John Benjamins. 31-73.

Mayer, Thomas & Michael Cysouw. 2014. Creating a Massively Parallel Bible Corpus. Proceedings of the International Conference on Language Resources and Evaluation (LREC), Reykjavik, 3158-3163. https://aclanthology.org/L14-1215/.

Mithun, Marianne. 1987. Is basic word order universal? In: Tomlin, R. S. (ed.), *Coherence and grounding in discourse: outcome of a symposium, Eugene, Oregon, June 1984 Typological studies in language.* Amsterdam, Philadelphia: John Benjamins. 281-328.

Nau, Nicole. 1996. Ein Beitrag zur Arealtypologie der Ostseeanrainersprachen. In: Boretzky, Norbert (ed.), *Areale, Kontakte, Dialekte, Sprachen und ihre Dynamik in mehrsprachigen Situationen*. [Bochum-Essener Beitrage zur Sprachwandelforschung, 24]. Bochum: Brockmeyer.

Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time.* London / Chicago: University of Chicago Press.

Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P, Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette M, Ribeiro Cunha E, Smith T, Stier A, Ter Braak C, Weedon J, Borman T. 2025. _vegan: Community Ecology Package_. R package version 2.6-10, <https://CRAN.R-project.org/package=vegan>.

Plungian, Vladimir A. 2023. The parallel corpus as a grammar database and the New Testament as a parallel corpus (Preface). In: Khomchenkova, Irina A., M. D. Vojejkova, N. M. Zaika, M. L. Kisilier, G. A. Mol'kov & A. Ju. Urmanchieva (eds.), *Studies in the theory of grammar, issue 9. The parallel corpus as a grammar database and the New Testament as a parallel corpus.* Acta Linguistica Petropolitana. Transactions of the Institute for Linguistic Studies. Vol. 19 part 3. 15-38.

Pozharickaja, Sofia K. 2011. On the Areal Distribution of Participial Forms in Russian Dialects. In: Seržant, Ilja A. & Björn Wiemer (eds.), *Contemporary Approaches to Dialectology: The area of North, Northwest Russian and Belarusian vernaculars*. Bergen: John Grieg AS. 109-129.

Ranacher, Peter, Nico Neureiter, Rik van Gijn, Barbara Sonnenhauser, Anastasia Escher, Robert Weibel, Pieter Muysken & Balthasar Bickel. 2021. Contact-tracing in cultural evolution: a Bayesian mixture model to detect geographic areas of language contact. *Journal of The Royal Society Interface. Royal Society* 18(181). 20201031.

Ross, Malcolm D. 2007. Calquing and metatypy. *Journal of Language Contact* 1(1). 116–143.

Seifart, Frank. 2015. Does structural-typological similarity affect borrowability? *Language Dynamics and Change* 5(1). 92–113.

Selting, Margret & Elizabeth Couper-Kuhlen 2000. Argumente für die Entwicklung einer 'interaktionalen Linguistik 1. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 1, 76-95. (www.gespraechsforschung-ozs.de)

Seržant, Ilja A.. 2015. Dative experiencer constructions as a Circum-Baltic isogloss. In: P. Arkadiev, A. Holvoet, B. Wiemer (eds.), *Contemporary Approaches to Baltic Linguistics*. De Gruyter. 325-348.

Seržant, Ilja A.. 2016. External possession and constructions that may have it. *Sprachtypologie und Universalienforschung STUF* 69(1). 131-169.

Seržant, I. 2021. Slavic morphosyntax is primarily determined by the geographic location and contact configuration. Scando-Slavica 67(1), 65-90.

Seržant, Ilja A., aop. Circum-Baltic Convergence Area. Greenberg, Marc (ed.), *Encyclopedia of Slavic Languages and Linguistics Online*. Brill.

Seržant, Ilja A., forthc. Statistical signal vs. areal/universal/genealogical pressure: Commentary on “Replication and methodological robustness in quantitative typology” by Becker and Guzmán Naranjo. *Linguistic Typology.* Accepted.

Seržant, Ilja A., Björn Wiemer, Eleni Bužarovska, Martina Ivanová, Maxim Makartsev, Stefan Savić, Dmitri Sitchinava, Karolína Skwarska, Mladen Uhlik. 2022. Areal and diachronic trends in argument flagging across Slavic. In: Eystein Dahl (ed.), *Alignment and Alignment Change in the Indo-European Family*. Oxford: OUP. 300-327.

Seržant, Ilja A., Daria Alfimova, Petr Biskup, Ivan Seržants. Forthc. Efficient sentence processing significantly affects the position of objects in Russian. *Linguistics.*

Siewierska, Anna & Uhlířová, Ludmila. 1998. An overview of word order in Slavic languages. In: Anna Siewierska (ed.), Constituent Order in the Languages of Europe. Berlin, New York: De Gruyter Mouton. 105-150.

Sinnemäki, Kaius, Francesca Di Garbo, Ricardo Napoleão de Souza & T. Mark Ellison. 2024. A typological approach to language change in contact situations. *Diachronica*. Ahead of print.

Stolz, Thomas 1991. *Sprachbund im Baltikum? Estnisch und Lettisch im Zentrum einer sprachlichen Konvergenzlandschaft.* [Bochum-Essener Beiträge zur Sprachwandelforschung, 13]. Bochum: Brockmeyer.

Tanaka, Hiroko. 2005. Grammar and the “timing” of social action: Word order and preference organization in *Japanese Language in Society* 34, 389–430.

Tosco, M. 2000. Ts there an Ethiopian language area? *Anthropological Linguistics* 42. 329-365.

Trubetzkoy, Nikolai S. 1923. Vavilonskaja bašnja i smešenie jazykov. *Evrazijskij vremennik* 3. 107–124.

Trubinskij, V. I. 1984. *Očerki russkogo dialektnolo sintaksisa.* Leningrad: Izdatel’stvo Leningradskogo universiteta.

Trubetzkoy, Nikolai S. (1928): [Proposition 16]. *Acts of the 1st International Congress of Linguistics 17-18.* Leiden.

Van den Heuvel, Wilco. 2020. Romani Bible Translation and the Use of Romani in Religious Contexts. In: Beníšek, Michael. 2020. The historical origins of Romani. In Yaron Matras and Anton Tenser (eds.), The Palgrave Handbook of Romani Language and Linguistics, 459-486. London: Palgrave Macmillan.

Verhagen, Arie. 2005. *Constructions of Intersubjectivity. Discourse, Syntax, and Cognition.* Oxford: OUP.

Vihman, Virve-Anneli & George Walkden. 2021. Verb-second in spoken and written Estonian. *Glossa: a journal of general linguistics* 6(1): 15. 1–23.

Watkins, Calvert.1963. Preliminaries to a historical and comparative analysis of the syntax of the Old Irish verb. Celtica 6. 1–49.

Wichmann, Søren. 2010. Internal language classification. In Vit Bubenik and Silvia Luraghi (eds.), *The Continuum Companion to Historical Linguistics*. New York: Continuum International. 70-86.

Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2024. Glottolog 5.1. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.14006617 (Available online at http://glottolog.org, Accessed on 2025-02-12.)

Slowikowski K (2024). _ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'_. R package version 0.9.6, <https://CRAN.R-project.org/package=ggrepel>.

South A, Michael S, Massicotte P (2024). _rnaturalearthdata: World Vector Map Data from Natural Earth Used in 'rnaturalearth'_. R package version 1.0.0, <https://CRAN.R-project.org/package=rnaturalearthdata>.

Massicotte P, South A (2023). _rnaturalearth: World Map Data from Natural Earth_. R package version 1.0.1, <https://CRAN.R-project.org/package=rnaturalearth>.
Pebesma, E., & Bivand, R. (2023). Spatial Data Science: With Applications in R. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016
Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009
H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
Moroz G (2017). _lingtypology: easy mapping for Linguistic Typology_.<https://CRAN.R-project.org/package=lingtypology>.
Kassambara A (2023). _rstatix: Pipe-Friendly Framework for Basic Statistical Tests_. R package version 0.7.2, <https://CRAN.R-project.org/package=rstatix>.