Subjects tend to be coded only once: Corpus-based and grammar-based evidence for an efficiency-driven trade-off

Aleksandrs Berdicevskis
Språkbanken
University of Gothenburg
aleksandrs.berdicevskis@gu.se

Karsten Schmidtke-Bode
Department of English and American Studies
Friedrich Schiller University Jena
karsten.schmidtkebode@uni-jena.de

Ilja Seržant
Project "Grammatical Universals"
Leipzig University
ilja.serzants@uni-leipzig.de

Abstract

Using data from the *World Atlas of Language Structures* and the *Universal Dependencies* treebanks, we provide converging evidence from linguistic typology and comparative corpus linguistics for an efficiency-based trade-off in the encoding of referentially accessible subjects. Specifically, when familiar subjects are marked as bound elements attaching to the verb, the chances of having obligatory independent subject pronouns decrease significantly across the world's languages. At the same time, there is a trend against not encoding the subject at all, leading us to postulate an overall tendency to encode familiar subjects once and only once in a neutral topic-comment utterance. This tendency is mirrored in more fine-grained corpus data from Slavic: East Slavic languages, in contrast to the other members of the genus, have past forms without verbal subject encoding, and it is precisely with these (former participle) forms that the use of independent subject pronouns is significantly higher than with other, non-participial verb forms. By contrast, the occurrence of independent subject pronouns does not differ across various verb forms in other Slavic languages, as none of them has been affected by a loss of verbal subject encoding.

1 Introduction and background

Trade-offs are a particular type of cross-linguistic tendency, which can usually be described as "in languages where feature X is strongly expressed, feature Y tends to be absent or weakly expressed, and vice versa". Examples of trade-offs include, for instance, the following: rich morphological marking of core arguments is negatively correlated with rigidity of word order (Sinnemäki, 2014; Futrell et al., 2015; Levshina, 2019), and, more generally, rigidity of word order is negatively correlated with rich word structure (Koplenig et al., 2017); paradigm size is negatively correlated with irregularity (Cotterell et al., 2019); word length is negatively correlated with phonotactic complexity (Pimentel et al., 2020).

Trade-offs can be explained as adaptations to communicative efficiency (Gibson et al., 2019) and/or learnability (Kirby et al., 2015): some overt coding is necessary to convey information robustly, but redundancy is undesirable (Berdicevskis and Eckhoff, 2016; Fedzechkina et al., 2017). Thus, the identification and description of trade-offs is important for the ongoing discussion about the extent to which language structure is shaped by adaptive pressures (Schmidtke-Bode et al., 2019). However, even the current wealth of databases, corpora and computational tools does not make the identification of cross-linguistic generalizations a trivial task, in large part due to the danger of spurious correlations (Ladd et al., 2015). The most reliable way of identifying a cross-linguistic generalization is to demonstrate its presence through different approaches that rely on different kinds of data (Roberts, 2018), such as typological surveys, corpus-based studies, psycholinguistic experiments, diachronic investigations and computational modelling (Bickel et al., 2015; Blasi et al., 2019; Bentz and Berdicevskis, 2016).

In this paper, we take such an approach in order to adduce new empirical evidence for a conspicuous trade-off that has long been discussed in the typological literature, viz. the trade-off in the encoding

of the subject. Subjects, especially those of transitive clauses, generally tend to be highly accessible referents (Du Bois, 2003; Du Bois, 1987; Siewierska, 2004). They are thus less likely to be encoded by full nominal phrases (NPs) than by *reduced referential devices* (Kibrik, 2011) such as independent pronouns and indexation (Ariel, 1990). We use *indexation* as an umbrella term for subject markers that are phonologically bound to the verb (i.e. verbal affixes and clitics) and index referential features of the subject (Haspelmath, 2013), typically person and number, but possibly also gender.

These different types of reduced referential devices are illustrated by the following examples:

(1) Norwegian Bokmål (independent subject pronoun)
Han sov.
'He slept.'

(2) Chalcatongo Mixtec (subject clitic):
Ni-éé=rí *staà.*[1]
CMPL-eat=1SG tortilla
'I ate.'

(3) Spanish (verbal affix):
ve-o
see-1SG.PRS
'I see'

Although languages vary as to the type of reduced referential device they conventionally employ, we argue that there is a strong trade-off pressure that constrains the typological distribution of the devices.

In particular, the trade-off is such that languages disprefer (a) doubling of reduced referential devices to encode the subject and (b) not encoding the subject at all. In other words, if — in information-structurally neutral clauses — a subject is encoded by a reduced referential device, there is a certain functional pressure to encode it once and only once in the clause.

This can be seen in the examples above, where the independent subject pronoun in (1) occurs with a verb that is itself unmarked for the subject, and the opposite situation holds in (2) and (3). What should be dispreferred and hence cross-linguistically rare according to our trade-off hypothesis is languages like German, which marks accessible subjects twice:

(4) German:
Wir lauf-en *schnell.*
we run.PRS-1PL fast
'We run/are running fast.'

Furthermore, the trade-off also predicts that the option of not encoding the subject at all, as in Chinese, is dispreferred, as it potentially engenders ambiguities in the unfolding discourse and thus requires additional processing effort ("hidden complexity" in Bisang (2015)) and risks a less accurate transfer of information. Although languages differ substantially in the degree to which they tolerate the omission of accessible referents in discourse (Bickel, 2003), our trade-off hypothesis leads us to assume that there is a cross-linguistic tendency against not encoding the subject, especially outside of closely tied syntactic units such as coordinate and subordinate clauses (Siewierska, 2004, p.22).Thus, despite the fact that the choice of the particular reduced referential device is subject to cross-linguistic variation, we suppose that languages tend to converge on optimizing the patterns by avoiding both redundancy (double encoding) and potential ambiguity (no encoding).

The complementary nature of independent subject pronouns and subject indexes has been a prominent feature in the discussion of the "null subject" (or "pro-drop") "parameter" in the formal-generative literature, where it was originally assumed that "only languages with rich verb agreement can license null subjects" (Taraldsen's (1980) generalization, as summarized in D'Alessandro (2015, p.219)). While

[1] CMPL = completive; example from Macaulay (1996, p. 141)

this generalization has been differentiated and refined by subsequent research (e.g. Rizzi (1982); Huang (1984); Müller (2006); Nicolis (2005); Roberts (2009)), its underlying premise is usually that the optionality of subject pronouns is viewed as a variable feature of innate linguistic representations ("Universal Grammar"). In the present paper, by contrast, we see it as a usage-based phenomenon that results from the strive for efficient communication.

Against this background, we probe the indexation vs pronoun trade-off by two complementary approaches. First, using the *World Atlas of Language Structures online* (Dryer and Haspelmath, 2013), we conduct a typological analysis on a broad sample of languages (Section 2). The survey shows that there is indeed a correlation between the presence of indexation, on the one hand, and the optionality of independent pronouns, on the other, and vice versa. This global correlation, per se, however, is not enough to establish a causal link between the two grammatical phenomena.

In a second study (Section 3), we thus perform a more specific, finer-grained corpus study on Slavic languages, using *Universal Dependencies* treebanks (Zeman et al., 2020). After all, in actual discourse, the two phenomena are not binary, but gradual: the proportion of sentences without an independent pronominal subject or in which the subject is indexed on the verb can vary greatly, so that the potential values are not limited to 0 and 1. Corpus-based approaches allow us to capture this variation (Levshina, 2019). We show that there is a split between East Slavic languages, on the one hand, and other modern Slavic languages, on the other. East Slavic languages have a number of constructions (most saliently, past tense) where the verb form (historically a participle) does not allow indexation. We show that in East Slavic, independent pronominal subjects are more frequent in these constructions than in those where indexation is possible, and that they are more frequent in East Slavic than in other Slavic languages (both in participial past-tense constructions and overall).

The results of the two studies, combined with previous quantitative work on other languages, provide strong evidence in favour of our hypothesized trade-off. Before we begin with the analyses, it needs to be pointed out that our prediction is limited to reduced referential devices; we do not make any predictions about full referential devices (NPs), since these fulfill a very different function in discourse.

2 Typological evidence

Our typological approach is similar in spirit to Gilligan's (1987) seminal investigation in being based on a sample of the world's languages, but it draws on a much larger and more contemporary database as well as completely different analytical tools. Specifically, we use a sample of of 241 languages for which data from two *WALS* chapters are available simultaneously, namely Dryer (2013) and Siewierska (2013). Dryer surveys the preferred expression of pronominal subjects, while Siewierska's chapter is concerned with the presence of verbal person marking. These surveys follow intricate coding schemes that come with the usual challenges of (i) reducing the variation range of human languages to a handful of types (see, e.g., Holmberg (2017) for a critique of some of Dryer's categories) and (ii) coding grammatical features whose presence of absence is variable rather than consistent in many languages (e.g. differential indexation). On top of that, the two coding schemes need to be both harmonized and further reduced for our purposes, as we are not interested in all the different types of subject expression and person marking, but in the general pattern of their interaction. Being aware of the risks that are harboured by such further interference with the data, we have opted to recode the two data sets as described in Table 1.

The decision to treat Dryer's "subject clitics" in the same fashion as his subject affixes on verbs is in line with our earlier definition of *indexation*, and this is also reflected by the fact that Siewierska analyzes them as verbal person markers (as long as one of their potential host words is actually the verb). Note that we left out Dryer's "mixed" category.

The 241 languages in this sample come from all six macro areas distinguished in *WALS*, spanning 100 language families (e.g. "Indo-European") and 179 lower genetic groupings called genera (e.g. "Germanic"). Since we thus have multiple data points for at least some of the language families and genera

[2]We understand Dryers category "subject pronouns in other position" to mean that these pronouns are obligatory. Note that this category in Dryer's coding may occasionally have been taken to instantiate subject clitics by Siewierska. For this reason, we also run an alternative model of the data in which this category is removed from the analysis (see footnote 4 below).

Feature	**Variable name in our model**	**Our coding scheme**	**Conflates the original *WALS* categories of ...**
Expression of pronominal subjects (Dryer, 2013)	Subject pronouns	Obligatory	obligatory pronouns in subject position, subject pronouns in other positions[2]
		Optional	subject affixes on verb, subject clitics, optional pronouns
Verbal person marking (Siewierska, 2013)	Subject indexation	Present	A&P, A only, A or P
		Absent	P only, No verbal person marking

Table 1: Coding of the typological data

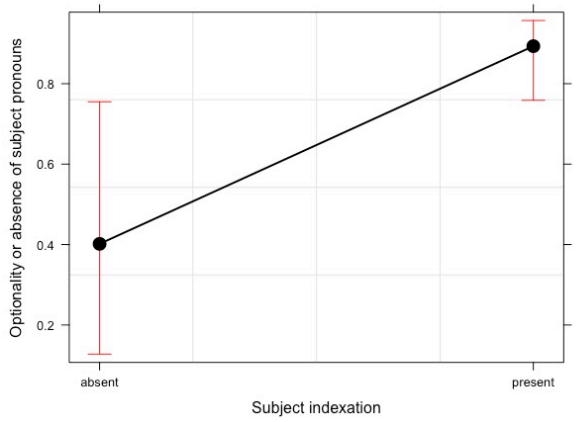

Figure 1: Visual representation of the effect of indexation on the optionality of subject pronouns in our model

in the sample, we use mixed-effects logistic regression modelling to take these dependencies into account. Specifically, we model the probability of having optional (or no) subject pronouns as a function of indexation, but control for repeated measurements. First, by modelling random intercepts for FAMILY and GENUS, we allow the genealogical units to have different baseline preferences for independent subject pronouns. Second, following arguments by Dryer (1989) and Bickel (2013), we assume that the most robust signals for universals come from those effects which take the same directionality across all of the world's macro areas. For this reason, we also include in our model by-AREA random slopes for the hypothesized effect of person marking on the occurrence of subject pronouns. We thus arrive at the following model formula[3]:

$$PronSubj \sim Index + (1|Family) + (1|Genus) + (1 + Index|Area)$$

The results of the modelling process show that the fixed effect of subject indexation on the absence of obligatory subject pronouns is significant ($\beta = 2.52$, $z = 2.3$, $p = 0.021$): on average, the odds for optional subject pronouns ("pro-drop") increase by 12.47 when we go from "absent" to "present" verbal person marking (Figure 1, which shows actual probabilities rather than odds).

[3]All statistical analyses were performed in *R* 3.3.0 (R Core Team, 2016), using the packages `lme4` (Bates et al., 2015), `effects` (Fox and Hong, 2009) and `rms` (Harrell Jr, 2020).

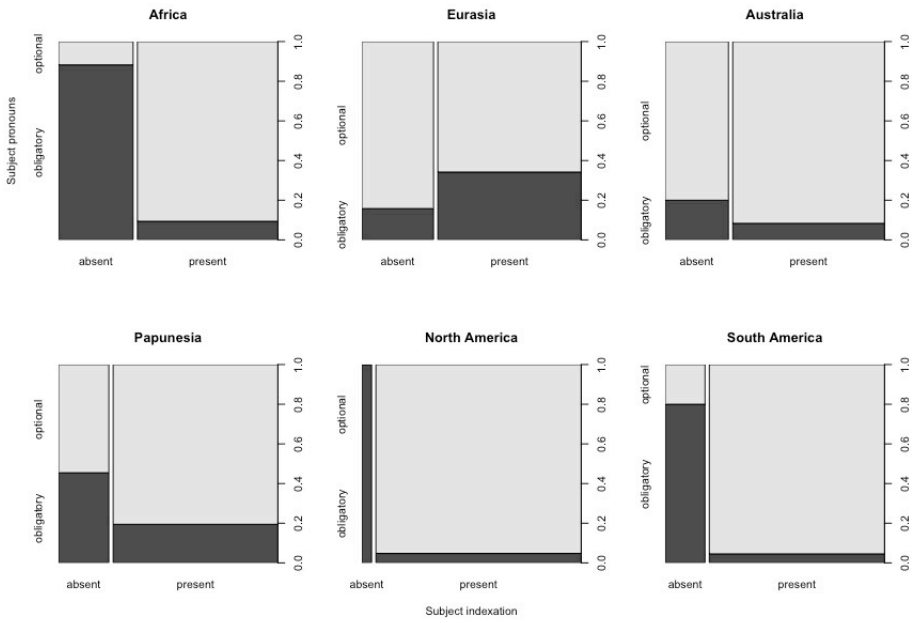

Figure 2: Areal distribution of the effect of indexation on the optionality of subject pronouns in the *raw* data (i.e. uncontrolled for genealogical dependencies)

Our model shows no traces of overdispersion and achieves very good discrimination (C = 0.88, Somers' D_{xy} = 0.75). In this respect, there appears to be robust cross-linguistic evidence for a trade-off between independent subject pronouns and subject indexation, indeed.

However, it must be conceded that our model is considerably "stressed" when it comes to the random effects structure relating to AREA, where it shows large standard deviations for both the intercept (sd = 1.56) and the slope adjustments (sd = 2.33). If we inspect the raw data again for the different macro areas (Figure 2), we can see that this problem mainly stems from the fact that one large and diverse area – Eurasia – goes against the otherwise consistent cross-linguistic trend, i.e. it reverses our hypothesized effect of indexation on subject pronouns. This is due to languages like French, German or Georgian on the one hand, which have obligatory subject pronouns despite verbal person marking (see Seržant (forthcoming a; forthcoming b) about this phenomenon), and to certain North and particularly South East Asian languages (e.g. Nivkh, Japanese, Korean, Chinese, Vietnamese, Burmese) which allow optional subject pronouns despite the lack of verbal person marking (for the latter, see e.g. Bisang (2014) for a diachronic account).[4]

The typological data also provide evidence for a general trend against having the option of not encoding the subject at all. We can illustrate this clearly again by juxtaposing how often the combination "no indexation and optional subject pronouns" (= no encoding occurs in discourse) is found with all other combinations where some subject encoding is always present. As can be seen in Table 2, the language families with at least one member of the no-encoding type are in the clear minority in all macro areas except again for Eurasia.

In conclusion, there is significant but not entirely consistent typological evidence for both aspects of the trade-off concerned here. In the corpus study to follow, we show, however, that our hypothesized pressures on subject encoding can be reliably detected even within Eurasia, and even in a genus which is

[4]If we run a mixed-effects regression model solely on the Eurasian languages in our sample (with random intercepts for FAMILY and GENUS), there is indeed an effect in the opposite direction from our trade-off hypothesis (i.e. the odds for obligatory subject pronouns decrease with the presence of verbal indexes), but it is not significant (p = 0.21 if the model does not contain by-FAMILY random slopes for the effect, and p = 0.33 if it does). If we run the same model as above but without Dryer's category "subject pronouns in other position", the model actually improves in the sense that the effect of indexation becomes even more robust (β = 4.112, z = 2.45, p = 0.014), as the distribution is even clearer in some of the macro areas and less pronouncedly reversed in Eurasia.

Macro area	No. of sample languages with no encoding	No. of sample families with no encoding
Africa	2/49	1/12
Eurasia	16/41	8/20
Australia	4/13	2/12
Papunesia	6/41	3/11
North America	0/44	0/27
South America	1/26	1/21

Table 2: Sample languages and families with no encoding of subject

mixed with regard to the possibility of not using independent subject pronouns. Specifically, we turn to the Slavic genus to buttress this claim.

3 Corpus-based evidence from Slavic

Indexation in modern East Slavic languages (Russian, Ukrainian, Belarusian) is different in past and non-past tenses. In the non-past, verb forms fusionally mark the subject's person and number; in the past, they mark number and, in the singular, also gender, but not person. The reason is that the modern past tense was originally the analytic perfect, consisting of a so-called *l*-participle and an optional copula. In East Slavic languages, the copula was gradually lost, and the participle was reanalyzed as a finite form. In all other Slavic languages, the copula was retained (in Polish, it has become a bound morpheme on the main verb), and it unambiguously marks the person of the subject.

If the trade-off described in Section 1 exists, we would expect that, in modern East Slavic languages, a subject would more often be encoded by an independent pronoun when the verb is in the past tense (or any other construction that is based on the *l*-participle, e.g. conditional) than in the tense-mood combinations that index person. Note that our decision to treat non-copular *l*-participle-based constructions as non-indexing is a simplification. While they do not index person, they do index number and gender, which can also aid the hearer's interpretation of subject reference. Nonetheless, it seems reasonable to believe that person marking provides much more information, and thus its absence should yield some effect, even if mitigated by the presence of number and gender.

In other Slavic languages, where person is always marked, either on an obligatory copula or a clitic, there should be no such difference between the *l*-participle and other tense-mood combinations. On the other hand, we may expect that in East Slavic, subjects of the verbs in *l*-participle-based constructions will be more often encoded by independent pronouns than in West and South Slavic.

Thus, we are effectively testing whether a typological generalization holds *within* individual languages. This enables us to perform a more direct test of the following causal relationship in East Slavic: the loss of indexation leads to the more frequent use of overt pronominal subjects. This is in line with what has been hypothesized about Slavic by Kibrik (2011). An opposing view had also been expressed in previous work (Ivanov, 1982; Zaliznjak, 2004), namely that subject pronouns expanded beyond their original limited usage, making person-marked copulas redundant and causing their gradual elimination. Note that under the latter view, however, there is no reason to expect differences either between *l*-participle and other forms in East Slavic or between East Slavic and other subgroups (see more in Section 4).

To test our prediction, we take all Slavic treebanks that are available in *Universal Dependencies* (UD) 2.6 (Zeman et al., 2020): Russian, Ukrainian, Belarusian (East); Czech, Polish, Slovak (West)[5]; Bulgarian, Slovenian, Serbian and Croatian (South). When several treebanks are available for one language, we concatenate them all. The resulting treebank sizes are very different across languages (from 13K tokens for Belarusian to 2.3M for Czech), but that is not a problem for our approach.

We take all finite verbs whose lemmas occur at least once in an *l*-participle-based construction (e.g.

[5]We include Upper Sorbian as well, but the treebank is too small and yields no datapoints that pass our filters.

	rus(E)	bel(E)	ukr(E)	pol(W)	cze(W)	slk(W)	crt(S)	srb(S)	blg(S)	slv(S)
l-participles	0.59	0.52	0.61	0.96	0.97	0.97	0.95	0.86	0.81	0.96
other forms	0.67	0.82	0.73	0.93	0.97	0.94	0.95	0.96	0.88	0.96

Table 3: Proportion of clauses without a free pronominal subject across **E**ast, **W**est and **S**outh Slavic.

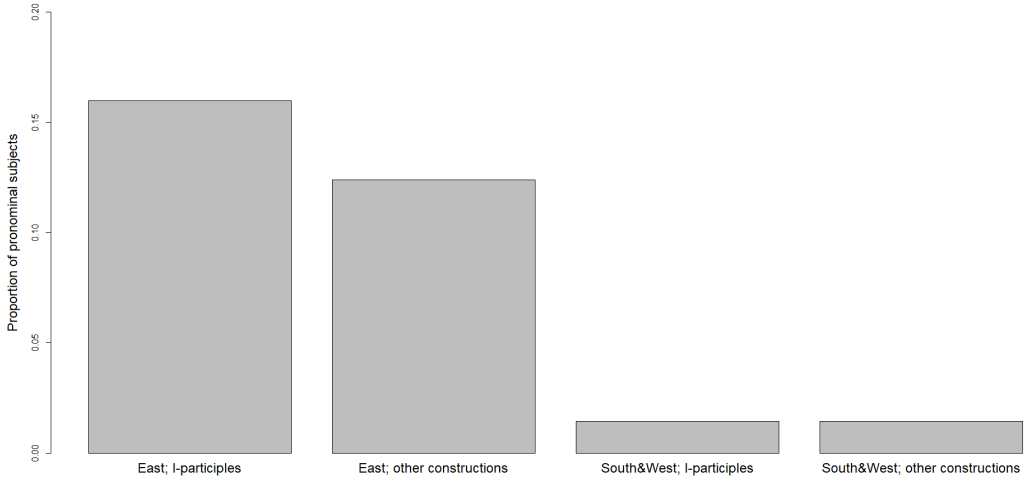

Figure 3: Proportion of clauses with an independent pronominal subject in the *raw* data (i.e. uncontrolled for language-specific and verb-specific effects)

past tense in East Slavic) and at least once in any other tense-mood combination. This filter is meant to exclude lemmas that occur only in a particular form. In addition, we require that the lemma occurs at least once with an independent nominal subject (related to the verb via NSUBJ). Here, we do not impose any additional limitations on the part-of-speech of the subject (it can be pronoun, noun or adjective), since the sole purpose of this filter is to exclude impersonal verbs (which never take a subject). For every verb token that passes these filters, we note whether it has an independent subject encoded by a personal pronoun or not. Since our trade-off hypothesis is limited to reduced referential devices, we do not include subjects encoded by NPs or anything else apart from personal pronouns (such clauses are ignored).

It could be argued that the analysis has to be narrowed down to first- and second-person pronouns, since in these cases the choice the speaker makes is clearly between a pronoun and its absence, while in the third person the choice is between an NP, a pronoun and absence of both. Furthermore, first- and second-person pronouns denote the referent much less ambiguously than the third-person pronouns. However, there is no way to automatically determine person in those East Slavic clauses where the verb is in an *l*-participle-based construction and the subject is not encoded by an independent pronoun (i.e. the person is not marked anywhere). For this reason, we do not choose the "first- and second-person only" design.

Our search for subjects does not extend beyond the clause boundaries. If the clause is coordinate or subordinate, we treat it irrespectively of what happens in other conjuncts or the main clause. In other words, in *She sings and walks* the verb *sing* would be treated as having a pronominal subject while the verb *walk* would not. The same is true for **She sings when walks* (which is a possible construction in Slavic).

We do not perform any analysis of the individual languages, but for illustrative purposes we provide information on the proportion of clauses without free prononimal subjects in different constructions across all languages in Table 3. Note that unlike all other languages, Bulgarian and Serbian do not follow the prediction, behaving rather like East Slavic languages.

The proportion of clauses with an independent pronominal subject across language groups and con-

Predictor	Estimate	SE	z value	p value
(Intercept)	-0.62	0.27	-2.3	0.020*
constr=person-marking	-0.38	0.03	-14.5	<0.001*
group=West&South	-2.58	0.32	-8.1	<0.001*
constr=person-marking x group=West&South	0.44	0.04	10.2	<0.001*

Table 4: Summary of the logistic-regression model: presence of a pronominal subject as predicted by construction and language group with by-VERB and by-LANGUAGE random effects. Asterisks denote significance at the 0.05 level.

struction types is visualized in Figure 3 (see also Table 3). The observed differences are in line with our expectations. To test whether they are significant, we fit a mixed-effects logistic regression model with the binary dependent variable being whether the subject is encoded by an independent pronoun. The predictors are CONSTRUCTION, or tense-mood combination (*l*-participle vs person-marked), GROUP (East vs West&South), and their interaction. We add by-VERB and by-LANGUAGE random intercepts in order to control for language-specific idiosyncrasies and individual lexical preferences of verbs. In *lme4* (Bates et al., 2015) notation, the model looks as follows:

$$PronSubj \sim construction * group + (1|verb) + (1|language)$$

The coefficient for CONSTRUCTION shows how frequently we find pronominal encoding of the subject in East Slavic in person-marking constructions (as opposed to *l*-participles), and we expect it to be negative. The coefficient for GROUP is meant to capture how frequently we find pronominal encoding of the subject in *l*-participle-based constructions in West and South Slavic (as opposed to East Slavic), and again, we expect it to be negative. We do not make specific predictions about the interaction, but we do not expect it to revert the individual effects.

We performed the calculations in *R* 4.0.2 (R Core Team, 2020), using the `lmerTest` package (Kuznetsova et al., 2017) to calculate *p*-values and `Hmisc` (Harrell Jr and others, 2020) to estimate discrimination. The total number of observations is 138,879; the number of unique verb lemmas is 9,363. The summary of the model is presented in Table 4.

It can be seen that all of our predictions are borne out. Within East Slavic languages, independent pronominal subjects are significantly more frequent in *l*-participle-based constructions. Within *l*-participle-based constructions, independent pronominal subjects are significantly more frequent in East Slavic languages than in other groups (cf. Seo (2001)). The interaction coefficient is positive, indicating that the joint effect is smaller than could be expected from individual coefficients. However, the interaction coefficient is noticeably smaller than the sum of the individual coefficients, suggesting that the joint effect is still significantly different from zero. The model shows very good discrimination (C = 0.88, Somers' D_{xy} = 0.76).

It is reasonable to expect that the means of encoding subject will strongly vary across clause types (simple sentence, subordinate, superordinate, coordinate), as we alluded to above. However, adding CLAUSE TYPE as a predictor leads to severe convergence problems, rendering the models unusable. Instead, we opt for a simpler way of controlling for a potential effect of CLAUSE TYPE. We fit a separate model with the same specification, but use only clauses from simple sentences (i.e. excluding all clauses from complex sentences: subordinate, superordinate and coordinate). The model yields similar results (see Supplementary materials).

Just like the typological data, the corpus data also provide evidence against not encoding subjects at all. We illustrate this by visualizing the number of clauses with no encoding, double encoding or one of two possible single encodings across all Slavic languages in our sample (Figure 4). Note also that "No encoding" column means "no person marking", while there still is gender and number marking.

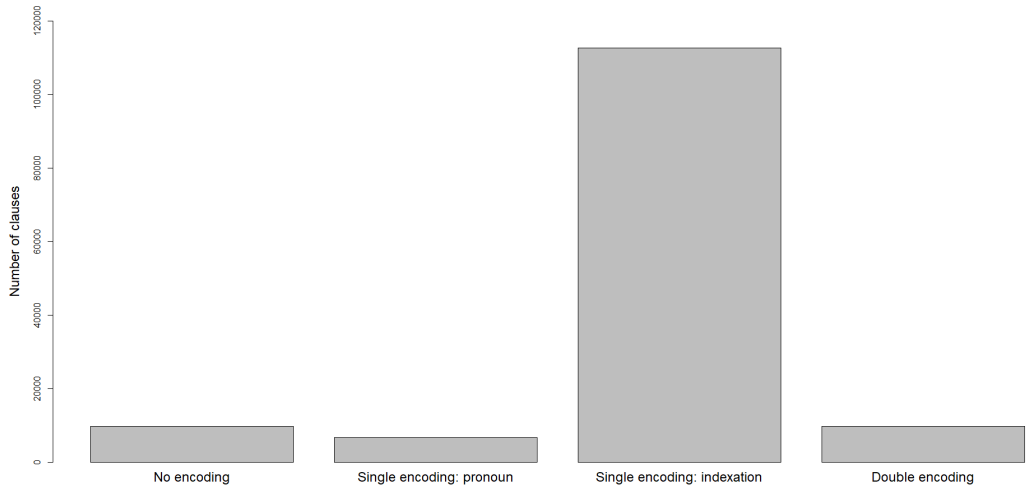

Figure 4: Distribution of subject-encoding strategies across all Slavic languages

4 Discussion

We have shown that grammar-based typological data and corpus-based data from UD provide converging evidence for our hypothesized trade-off between independent pronominal subjects and indexation. This, in turn, supports our claim that languages prefer to encode accessible subject referents once and only once within a clause. The typological study shows that the presence of subject indexation significantly increases the optionality of subject pronouns. The trend, however is not exceptionless: it is reversed in Eurasia, by an unusual confluence of double-encoding languages (e.g. German, Standard French, Georgian) and no-encoding languages (e.g. Nivkh, Japanese, Chinese). The UD study provides more direct evidence that the absence of indexation encourages speakers to encode accessible subject referents by independent pronouns significantly more often. It thus suggests, in keeping with the original *WALS* data, that there is also a certain pressure against not encoding the subject.

This observation raises an important question about the directionality of change: does the loss/emergence of indexation make pronouns more or less obligatory or is it the other way around (Kibrik, 2004)? Our evidence suggests that the causal path "loss of indexation $\rightarrow$ higher proportion of pronouns" is present in Slavic. Nonetheless, we cannot discard earlier claims to the opposite effect, viz. "higher proportion of pronouns $\rightarrow$ loss of indexation" (Ivanov, 1982; Zaliznjak, 2004), since we provide no evidence either for or against it. It may be that both causal paths exist (both within Slavic and universally).

Simonenko et al. (2019) addressed the question of directionality, performing a diachronic study on Medieval French. They show that the syncretization of verbal endings (which is presumably phonologically-driven and eventually leads to the near-disappearance of indexation) occurs at almost the same rates as the spread of pronominal subject encoding, which suggests that the two processes are likely to be related. They compare two models: in the first one, indexation and absence of independent subject pronouns are manifestations of the same grammatical property (it is unclear how this model can be interpreted outside the generative framework); in the second one, not encoding the subject is assumed to create a parsing difficulty and is thus dispreferred by language learners. The second model is shown to have a higher fitness, which is interpreted as evidence in favour of the causal link "loss of indexation $\rightarrow$ higher proportion of pronouns". That said, it should be noted that Simonenko et al.'s study is deeply rooted in generative assumptions, in particular, the constant rate hypothesis (Kroch, 1989), and is difficult to evaluate outside of this framework. Furthermore, the authors did not explicitly test the reverse causal path.

Our data support the earlier observation that pronominal subjects are more frequent in East Slavic than in other subgroups (Seo, 2001), not only with *l*-participle-based constructions, but in all clauses. This is in keeping with the common view that East Slavic languages are gradually undergoing a development

towards double encoding with both obligatory independent pronouns and verbal affixes, while most other modern Slavic languages remain to be single-encoding languages. It seems reasonable to assume that this is explained by analogical extension: pronominal subjects are spreading from *l*-participle-based constructions to other clauses (Kibrik, 2004).

A promising research avenue would be a quantitative diachronic investigation of the loss of the copula and the spread of the subject pronoun in Slavic. Our pilot study, however, suggests that the current treebanks of older stages of Slavic languages might not be large enough to yield robust results.

Note also that the trade-off is not absolute within Slavic (see Figure 4), as there are clauses where the subject is not encoded at all (*l*-participles without a pronoun in East Slavic languages) and clauses with double encoding (other forms with a pronoun). The non-negligible proportion of counterexamples both within and across languages suggests that the pressure for the optimization of subject encoding is relatively weak, and probably moderated by other factors. Double encoding of the subject, for instance, is normally reserved for the rare, pragmatically marked clause types (e.g. marked-focus or topic-shift subjects, cf. English *Me, I like booze* or *Him, he's crazy*[6]). Historically, such double encoding can spread to topic-comment clauses via the overuse of what originally used to be a pragmatically marked information structure, as described in detail in Givón (1976) (see also Ariel (2000)). This is, for example, what happened in the development from Old High German with its optional free pronouns into Modern German with obligatory free pronouns (Axel and Weiß, 2011). Double-encoding systems often turn into single-coding systems by abandoning the indexation, as in English or French (Siewierska, 2004, p.295). Absence of encoding, in turn, sometimes arises as a by-product of the emergence of new verbal forms based on nominalized structures which are not amenable to person marking themselves, such as Slavic *l*-participles. And as noted in Sections 1 and 2, typologically dispreferred structures may still spread locally to form areal phenomena, such as not encoding familiar subjects in the languages of South and Southeast mainland Asia (notably the Sinitic subfamily (Sino-Tibetan), the Mon-Khmer subfamily (Austroasiatic), Tai (Tai-Kadai), Hmong-Mien and Chamic (Austronesian), see Bisang (2014) for a historical account).

As was laid out at the beginning, we see the motivation behind the single-encoding pressure in the strive for efficient communication that equilibrates production effort and the robustness of information transfer (cf. also Jaeger and Buz (2017)). No encoding of a central discourse referent potentially jeopardizes the accurate transmission of messages. But double encoding is obviously redundant in pragmatically unmarked topic-comment clauses and is therefore costlier than necessary. In this respect, our findings confirm Haspelmath's form-frequency correspondence universal, according to which languages *generally* tend to have shorter forms for more frequent meanings (see Haspelmath's (2008a; 2008b) generalization across various earlier proposals for coding efficiency in the lexicon, e.g. Zipf (1935), and several domains of grammar (Greenberg (1966); Comrie (1989); Hawkins (2004))).

From this perspective, the frequent double encoding of pragmatically marked subjects (focal subjects, topic-shift subjects, etc.) is also explained: since pragmatically marked subjects are considerably less frequent in discourse than the pragmatically unmarked continuous-topic subjects (Givón, 1992), it is the pragmatically marked subjects that tend to select double encoding, which is costlier than the single encoding of topical subjects.

On a methodological level, our paper illustrates how coarse-grained but broad typological data and more fine-grained but narrower corpus data can fruitfully complement each other. From a technical point of view, it has become obvious to us, however, that the current UD annotation is still far from being fully harmonized. For instance, plural pronouns like 'we' are annotated as 'I'-PL in Slovenian and Upper Sorbian (probably due to the presence of dual forms), while in other languages, 'I' and 'we' are treated as separate lemmas. While such discrepancies are understandable and in certain ways beneficial, they may become a hindrance for cross-linguistic comparison, especially if not thoroughly documented.

The Supplementary materials, including scripts for extracting the data and running the statistical analysis, are openly available[7].

[6]Examples from Rodman (1997, p. 53).

[7]https://github.com/AleksandrsBerdicevskis/subject-encoding

Acknowledgements

KSB and IS were supported by European Research Council (ERC) Advanced Grant 670985, which is gratefully acknowledged.

References

Mira Ariel. 1990. *Accessing noun-phrase antecedents*. Routledge.

Mira Ariel. 2000. The development of person agreement markers: From pronouns to higher accessibility markers. In Michael Barlow and Suzanne Kemmer, editors, *Usage-based models of language*, pages 197–260. CSLI Publications.

Katrin Axel and Helmut Weiß. 2011. Pro-drop in the history of German from Old High German to the modern dialects. In Melani Wratil and Peter Gallmann, editors, *Null pronouns*, pages 21–52. John Benjamins.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1):1–48.

Christian Bentz and Aleksandrs Berdicevskis. 2016. Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 222–232, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Aleksandrs Berdicevskis and Hanne Eckhoff. 2016. Redundant features are less likely to survive: Empirical evidence from the Slavic languages. In S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, and T. Verhoef, editors, *The Evolution of language: Proceedings of the 11th International Conference (EVOLANGX11)*. Online at `http://evolang.org/neworleans/papers/85.html`.

Balthasar Bickel, Alena Witzlack-Makarevich, Kamal K. Choudhary, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky. 2015. The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PLOS ONE*, 10(8):1–22, 08.

Balthasar Bickel. 2003. Referential density in discourse and syntactic typology. *Language*, 79(4):708–736.

Balthasar Bickel. 2013. Distributional biases in language families. In Balthasar Bickel, Lenore A Grenoble, David A Peterson, and Alan Timberlake, editors, *Language typology and historical contingency*, pages 415–444. John Benjamins.

Walter Bisang. 2014. On the strength of morphological paradigms. In Martine Robbeets and Walter Bisang, editors, *Paradigm change: In the Transeurasian languages and beyond*, pages 23–60. John Benjamins.

Walter Bisang. 2015. Hidden complexity–the neglected side of complexity and its implications. *Linguistics Vanguard*, 1(1):177–187.

Damian Blasi, Steven Moran, Scott Moisik, Paul Widmer, Dan Dediu, and Balthasar Bickel. 2019. Human sound systems are shaped by post-neolithic changes in bite configuration. *Science*, 363(6432).

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago Press.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S Dryer. 1989. Large linguistic areas and language sampling. *Studies in Language*, 13(2):257–292.

Matthew S Dryer. 2013. Expression of pronominal subjects. In Matthew S Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

John Du Bois. 1987. The discourse basis of ergativity. *Language*, 63(4):805–855.

John Du Bois. 2003. Argument structure: Grammar in use. In Lorraine Kumpf and William Ashby, editors, *Preferred argument structure: Grammar as architecture for function*, pages 11–60. John Benjamins.

Roberta dAlessandro. 2015. Null subjects. In Antonio Fábregas, Jaume Mateu, and Michael Putnam, editors, *Contemporary linguistic parameters*, pages 201–226. Bloomsbury London.

Maryia Fedzechkina, Elissa L. Newport, and T Florian Jaeger. 2017. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive Science*, 41(2):416–446.

John Fox and Jangman Hong. 2009. Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software*, 32(1):1–24.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the third international conference on dependency linguistics (Depling 2015)*, pages 91–100.

Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in cognitive sciences*, 23(5):389–407.

Gary Gilligan. 1987. *A cross-linguistic approach to the pro-drop parameter*. Ph.D. thesis, University of California.

Talmy Givón. 1992. The grammar of referential coherence as mental processing instructions. *Linguistics*, 30(1):5–55.

Talmy Givón. 1976. Topic, pronoun, and grammatical agreement. In Charles N. Li, editor, *Subject and topic*, pages 149–188. Academic Press.

Joseph H Greenberg. 1966. *Language universals: With special reference to feature hierarchies*. Mouton.

Frank E Harrell Jr et al., 2020. *hmisc: Harrell Miscellaneous*. R package version 4.4-1.

Frank E Harrell Jr, 2020. *rms: Regression modeling strategies*. R package version 6.0-1.

Martin Haspelmath. 2008a. Creating economical morphosyntactic patterns in language change. In Jeff Good, editor, *Linguistic universals and language change*, pages 185–214. Oxford University Press.

Martin Haspelmath. 2008b. A frequentist explanation of some universals of reflexive marking. *Linguistic Discovery*, 6(1):40–63.

Martin Haspelmath. 2013. Argument indexing: a conceptual framework for the syntax of bound person forms. In Dik Bakker and Martin Haspelmath, editors, *Languages across boundaries: Studies in memory of Anna Siewierska*, pages 209–238. De Gruyter Mouton.

John A Hawkins. 2004. *Efficiency and complexity in grammars*. Oxford University Press.

Anders Holmberg. 2017. Linguistic typology. In Ian Roberts, editor, *The Oxford handbook of Universal Grammar*, pages 355–376. Oxford University Press.

C-T James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry*, pages 531–574.

Valerij V Ivanov. 1982. Istorija vremennyx form glagola. In Ruben Avanesov and Valerij Ivanov, editors, *Istoricheskaja grammatika russkogo jazyka. Morfologija. Glagol*, pages 25–131. Nauka.

T Florian Jaeger and Esteban Buz. 2017. Signal reduction and linguistic encoding. In Eva Fernández and Helen Smith Cairns, editors, *The Handbook of psycholinguistics*, pages 38–81. Wiley-Blackwell.

Andrej Kibrik. 2004. Zero anaphora vs. zero person marking in Slavic: A chicken/egg dilemma? In António Branco, Tony McEnery, and Ruslan Mitkov, editors, *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, pages 87–90. Edicoes Colibri.

Andrej Kibrik. 2011. *Reference in discourse*. Oxford University Press.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer. 2017. The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort. *PLOS ONE*, 12(3):1–25, 03.

Anthony S. Kroch. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1(3):199–244.

Alexandra Kuznetsova, Per B Brockhoff, Rune HB Christensen, et al. 2017. lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

D Robert Ladd, Seán G Roberts, and Dan Dediu. 2015. Correlational studies in typological and historical linguistics. *Annu. Rev. Linguist.*, 1(1):221–241.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.

Monica Ann Macaulay. 1996. *A grammar of Chalcatongo Mixtec*. Univ of California Press.

Gereon Müller. 2006. Pro-drop and impoverishment. In Patrick Brandt and Eric Fuss, editors, *Form, structure, and grammar. A festschrift presented to Günther Grewendorf on occasion of his 60th birthday*, pages 93–115. Narr.

Marco Nicolis. 2005. *On pro drop*. Ph.D. thesis, University of Siena.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

R Core Team, 2016. *R: A language and environment for statistical computing, Version 3.3.0*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team, 2020. *R: A language and environment for statistical computing, Version 4.0.2*. R Foundation for Statistical Computing, Vienna, Austria.

Luigi Rizzi. 1982. *Issues in Italian syntax*. Foris.

Ian Roberts. 2009. A deletion analysis of null subjects. In Theresa Biberauer, Anders Holmber, and Ian Roberts, editors, *Parametric variation: Null subjects in minimalist theory*, pages 58–87. Cambridge University Press.

Seán G. Roberts. 2018. Robust, causal, and incremental approaches to investigating linguistic adaptation. *Frontiers in Psychology*, 9:166.

Robert Rodman. 1997. On left dislocation. In Frans Zwarts Elena Anagnostopolou, Henk van Riemsdijk, editor, *Materials on left dislocation*, pages 31–54. John Benjamins.

Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis, and Ilja A Seržant, editors. 2019. *Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence*. Language Science Press.

Seunghyun Seo. 2001. *The frequency of null subject in Russian, Polish, Czech, Bulgarian and Serbo-Croatian: An analysis according to morphosyntactic environments.* Ph.D. thesis, Indiana University.

Ilja A Seržant. forthcoming-a. Cyclic changes in verbal indexes are not drift processes. *Folia Linguistica Historica*.

Ilja A Seržant. forthcoming-b. The dynamics of slavic morphosyntax is primarily determined by the geographic location and contact configuration. *Scando-Slavica*.

Anna Siewierska. 2004. *Person*. Cambridge University Press.

Anna Siewierska. 2013. Verbal person marking. In Matthew S Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Alexandra Simonenko, Benoît Crabbé, and Sophie Prevost. 2019. Agreement syncretisation and the loss of null subjects : quantificational models for medieval french. *Language Variation and Change*, 31(3):275–301.

Kaius Sinnemäki. 2014. Complexity trade-offs: A case study. In Frederick Newmeyer and Laurel Preston, editors, *Measuring grammatical complexity*, pages 179–201. Oxford University Press.

Knut Tarald Taraldsen. 1980. *On the nominative island condition, vacuous application and the that-trace filter*. Indiana Univ. Linguistics Club.

Andrej Zaliznjak. 2004. *Drevnenovgorodskij dialekt. 2-e izd., dop. i pererab*. Jazyki slavjanskoj kul'tury.

Daniel Zeman, Joakim Nivre, et al. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

George Kingsley Zipf. 1935. *The psychobiology of language*. Houghton-Mifflin.