Laura Becker, Matías Guzmán Naranjo and Samira Ochs

5 Socio-linguistic effects on conditional constructions: A quantitative typological study

Abstract: Recent typological studies have shown that socio-linguistic factors have a substantial effect on at least certain structures of language. However, we are still far from understanding how such factors should be operationalized and how they interact with other factors in shaping grammar. To address both questions, this study examines the influence of socio-linguistic factors on the number of dedicated conditional constructions in a sample of 374 languages. We test the number of speakers, the degree of multilingualism, the availability of a literature tradition, the use of writing, and the use of the language in the education system. At the same time, we control for genealogical, contact, and bibliographical biases. Our results suggest that the number of speakers is the most informative predictor. However, we find that the association between the number of speakers and the number of dedicated conditional constructions is much weaker than assumed, once genealogical and contact biases are controlled for.

Acknowledgments: We wish to thank the discussants from the 2021 SLE workshop "Integrating socio-linguistic and typological perspectives on language variation: methods and concepts", the participants of the Freiburg Linguistics reading group, Marvin Martiny, Uta Reinöhl and Maria Vollmer as well as two anonymous reviewers for their valuable comments on earlier versions of this study. Furthermore, this paper was supported by a Junior Fellowship from the Freiburg Institute for Advanced Studies (FRIAS), University of Freiburg (Germany), by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement 834050), and the Emmy Noether project 'Bayesian modelling of spatial typology' (project number 504155622).

Laura Becker, University of Freiburg, Belfortstraße 18, 79098 Freiburg im Breisgau,
e-mail: laura.becker@linguistik.uni-freiburg.de
Matías Guzmán Naranjo, University of Freiburg, Belfortstraße 18, 79098 Freiburg im Breisgau,
e-mail: matias.guzman.naranjo@linguistik.uni-freiburg.de
Samira Ochs, Leibniz-Institut für Deutsche Sprache (IDS), R5, 6-13, 68161 Mannheim,
e-mail: ochs@ids-mannheim.de

https://doi.org/10.1515/9783110781168-005

1 Introduction

Following the spirit of socio-typology of Trudgill (2008), a number of recent quantitative typological and crosslinguistic studies have shown that socio-linguistic factors such as the number of speakers or the proportion of L2-speakers have a substantial effect on at least certain structures of language (Bentz & Winter 2013; De Busser & LaPolla 2015; Karlsson, Miestamo & Sinnemäki 2008; Ladd, Roberts & Dediu 2015; Lupyan & Dale 2016; Sinnemäki 2020; Sinnemäki & Di Garbo 2018; Trudgill 2008, 2011b). The idea is that the structure of a linguistic community can have an effect on the grammatical properties that the language of that community may develop. The prime example of this is the claim that larger communities with many adult L2-speakers will tend to develop simpler morphology as adults have more difficulties with learning complex morphology (Trudgill 2011b). On the other hand, smaller communities with a high number of bilingual children can develop more complex morphological systems due to transfer (Trudgill 2010: 301–306).

While the question of how socio-linguistic factors can shape grammar is a promising research area, we are still far from having a solid understanding of which factors play a role in shaping grammar, how the different factors interact, or how they should be operationalised. So far, most large scale studies in socio-linguistic typology have focused on only one or two factors, namely population size and L2-speaker proportion. Besides these factors, the use of the language in writing and its literature tradition may also impact its grammatical properties over the course of time. Additionally, while most work on socio-linguistic typology has tried to control for genetic bias, there are other sources of bias that we need to consider as well: areal effects due to contact and diffusion, as well as cultural or socio-linguistic, typological and bibliographical biases. Despite the areal bias being well known, it is also harder to control or to account for, which explains that it has not received sufficient attention in typological studies in the past.

The linguistic phenomenon that we will examine in this study is the expression of conditionals. Conditional constructions lend themselves as a testing ground for the impact of socio-linguistic factors on language structure for two reasons. First, all languages have some way of expressing conditions (cf. Wierzbicka 1996: 68–70) and, second, conditional constructions show a large degree of crosslinguistic variation that we can make use of to determine the impact of socio-linguistic variables. In addition, previous work suggests that the degree of lexicalization, grammaticalization and explicitness of conditional constructions and markers is prone to be influenced by socio-linguistic factors (Martowicz 2011). Building on this, we compiled a sample of 374 languages and annotated for the number of dedicated conditional constructions in each language. We use this linguistic phenomenon to examine the

influence that genealogical, contact and socio-linguistic factors have on language once we consider their interaction.

This chapter is structured as follows: Section 2 provides a brief overview of conditional constructions and Section 3 introduces the known socio-linguistic factors that impact language structure. In Section 4, we present our case study, describing the sample and the annotation. Section 5 presents and analyzes our results, which are further discussed in Section 6. Section 7 concludes.

2 Conditional constructions: Preliminary remarks

Conditional expressions relate two events, a main event and a condition of that event. Formally, the condition can be expressed through an adverbial clause, which is also referred to as the protasis. The main event, often formally expressed in a main clause, is called the apodosis. Semantically, we can distinguish between three broad types of conditional expressions: real, hypothetical, and counterfactual conditionals (e.g. Hetterle 2015: 48–50, Kortmann 1997: 85, Thompson, Longacre & Hwang 2007: 255–256). Real conditionals refer to real present, past, future or general events. While the condition of the main event does not necessarily have to occur, once the condition occurs, the main event does so as well. Two examples of real conditionals from English are given in (1) and (2). Example (1) refers to a specific situation, and (2) contains a generic conditional, expressing a general truth.[1]

(1) *[If it's raining on my way home]$_{PRO}$, [I will get wet]$_{APO}$.*

(2) *[If you do not get enough sleep]$_{PRO}$, [you will be tired]$_{APO}$.*

In hypothetical conditionals, both the conditional and the main event are more imaginative and less likely to happen compared to real conditionals; they express what might be. In the case of example (3), the speaker expresses that the condition (meeting her friend), is unlikely in the first place, making the main event (that she does not recognise her) unlikely as well.

(3) *[If I met my friend from kindergarden]$_{PRO}$, [I would not recognise her]$_{APO}$.*

The third main type of conditional expressions is the counterfactual conditional. Counterfactuals express events that did not happen. In this case, it is presupposed

1 The protasis is marked by PRO and the apodosis by APO here and in the following examples.

that the condition was not met. An example of a counterfactual conditional is given in (4).

(4) *[If you had been in class today]$_{PRO}$, [you would have seen the new teacher]$_{APO}$.*

For the purposes of the present study, we include conditional constructions of all three types, i.e. real, hypothetical and counterfactual conditionals without further distinctions between the three types. We include all types because the distinction of different types is irrelevant for our research question, as we compare the effect of socio-linguistic factors on the overall number of dedicated conditional constructions in a given language.

We know about a few typologically common and less common properties of conditional constructions from the literature concerned with either adverbial clauses in general (Diessel & Gast 2012; Hetterle 2015; Kortmann 1997) or conditional expressions and constructions more specifically (Athanasiadou & Dirven 1997; Khrakovskij 2005; Podlesskaya 2001; Thompson, Longacre & Hwang 2007; Traugott et al. 1986). For instance, the statement containing the condition (protasis) usually precedes the statement of the main event (apodosis), as could be seen in examples (1) to (4). This order is also the default order across languages, although it can be reversed in certain cases depending on context and language-specific conditions.[2] For the purposes of the present study, the order of protasis and apodosis does not play a role in that we count pairs such as *if it rains, I will get wet* and *I will get wet if it rains* as a single construction.[3]

Another property concerns the type of conditional marker. Thompson, Longacre & Hwang (2007) note that an equivalent of the English *if* marker is typologically quite common, i.e. many languages use a subordinator of some sort in the protasis to express conditionality. Example (5) shows this for Goemai (Chadic, Nigeria). Goemai uses the marker *là* in the protasis to signal conditionality similarly to the use of *if* in English.

(5) ***[Là*** *góe=p'ét]$_{PRO}$* [*t'òng* *góe=múút*]$_{APO}$.
COND 2SG.M.S=exit.SG IRR 2SG.M.S=die. SG
'If you go out, you will die.'
Goemai (Hellwig 2011: 457)

2 Already Greenberg (1963: 66) proposes the order of protasis preceding the apodosis as being universally preferred. It is also this default order of the protasis preceding the apodosis that inspired analyses of conditional statements as topics (e.g. Haiman 1978; Podlesskaya 2001).

3 This is mainly a practical decision because most descriptions do not provide explicit information on the (preferred) order of the protasis and the apodosis outside of the examples shown.

Comrie (1986), Podlesskaya (2001) and Thompson, Longacre & Hwang (2007) observe that it is very common to mark the protasis as in English or Goemai, and that most languages do not use any obligatory marker in the apodosis. The constructions in our sample show the same trend. Some languages or single constructions in a given language may use a marker in the apodosis, but it is often used emphatically in addition to another marker in the protasis.[4] For instance, in their description of Yanyuwa (Pama-Nyungan, Australia), Kirton & Charlie (1996: 190–191) write: "[t]he apodosis is usually unmarked, but if the speaker wishes to emphasise the sureness of the consequence, then the apodosis is introduced by one of the following: *kulu* 'and, then', *mardalmarda barra* 'and, also' or *barra* 'then'." This is shown in (6) and (7) below. In (6), we see a conditional construction marked by *namba* in the protasis, and in (7), *barra* is used in addition to *namba* in the apodosis of the conditional expression. In such cases, we do however treat the expressions shown in (6) and (7) as two variants of a single construction, i.e. they are counted in as one. Languages almost always allow the (spontaneous) use of an additional marker equivalent to English *then* in the apodosis, but this is not always made explicit in the descriptions. This makes the consistent distinction of such variants very difficult crosslinguistically, and we therefore do not count them in as separate constructions.

(6) ***[Namba*** *kurdardi buyuka-wu]*$_{PRO}$ *yijini-nja-rra,* *wurnda* *ma-nja-rra*
if not fire-DAT kindle-PTCP-PRS WOOD break-PTCP-PRS
yijini-nja-rra-i, *baki* *wakara,* *buyuka,* *ji-walanyma-nji]*$_{APO}$.
kindle- PTCP-PRS-on.and.on and success fire it-emerge-PRS
"If (there is) no fire, (then there is) making fire (by twrling one firestick into another), breaking wood making fire on and on, and it's there! - fire! - it is coming."
Yanyuwa (Kirton & Charlie 1996: 176)

(7) ***[Namba*** *kari-wayka* *wabuda* *ki-walanyma-njima]*$_{PRO}$,
if from-down water it-emerge-POT
[barra *manthalmanthal* *nawu* *awara,* *wararr* *barra]*$_{APO}$.
then soft now ground mud now
'If the water should come up from down there, then the ground is soft, there is mud.'
Yanyuwa (Kirton & Charlie 1996: 191)

4 Out of 1142 conditional constructions in our dataset, 871 only use a marker in the protasis, 107 constructions have a marker in both clauses, 87 use no overt marker, 36 have an optional marker in the apodosis in addition to the one in the protasis, 13 feature a marker in the apodosis with an optional additional marker in the protasis, 23 only have a marker in the apodosis, and 5 constructions use the same marker either in the apodosis or in the protasis.

In addition, Thompson, Longacre & Hwang (2007: 256–257) mention that it is common to use a dedicated marker or construction in hypothetical and counterfactual conditionals, and less so in real conditionals, which are often expressed as temporal clauses. This is also what we find in our dataset; usually, if conditionality is expressed by juxtaposition only, we are dealing with a real conditional, in which case the main event can still occur. Hypothetical or counterfactual conditionals, on the other hand, are usually formally marked in some way. Examples (8) and (9) show this for Bengali (Indo-European, Bangladesh). Bengali has a conditional marker, *yôdi*, which is systematically used to mark counterfactual conditionals. This is shown in (8). Example (9) then shows that real conditional statements can be expressed by the juxtaposition of protasis and apodosis without the use of *yôdi*.[5] Because we will analyze the number of dedicated conditional constructions, expressions involving no dedicated marker such as the juxtapposition of two clauses in (9) will not count towards the number of constructions. In other words, a language that only marks conditional relations using the juxtapposition of clauses will be treated to have a count of 0 dedicated conditional constructions.

(8) ***[yôdi*** *ami* *susthô* *thaktam]*$_{PRO}$ *[tahôle* *côle* *yetam*
if I well be.PST.HAB.1 then move.PTCP go.PST.HAB.1
kothao]$_{APO}$.
somewhere
'If I were well, I would go away somewhere.'
Bengali (Thompson 2012: 243)

(9) *[bhorbœla* *sarṭer* *gɔlaŷ* *ṭai thake* *na]*$_{PRO}$, *[kœmôn*
dawn.hour shirt.GEN throat.LOC tie stay.PRS.3 not how
yœnô *khali* *ga* *mône* *hɔŷ* *tãr]*$_{APO}$.
as.if empty body mind.LOC be.PRS.3 he.HON.GEN
'If he does not have a tie round his neck by dawn, he feels somehow naked.'
Bengali (Thompson 2012: 246)

Although it is common to have a syntactic marker such as the subordinator *if* in English to signal the conditionality in the protasis, previous work has revealed much variation in how conditionals can be expressed (e.g. Khrakovskij 2005; Podlesskaya 2001; Thompson, Longacre & Hwang 2007). In our dataset, we find

5 It is likely that intonation and prosody play a role in those cases in which conditionals are expressed only by juxtapposition of two clauses, but a systematic analysis thereof, also for other languages, is not available yet.

various formal strategies as well. In a number of languages, conditionality is expressed morphologically by a verbal marker. This is shown for Oko (Benue-Congo, Nigeria) in (10). Other languages make use of nominal or nominalization strategies. For instance, in Kwini (Worroran, Australia) conditionals can be expressed through the use of the nominalizer *-ngay* which attaches to the verb in the protasis. As can be seen in (11), the nominalizer is the only formal marker of conditionality. Other languages use topic markers to express conditionals. One such example is shown in (12) from Shiwiar (Chicham, Ecuador). Here the topic marker *=ka* is used in the protasis to encode conditionality. Yet another strategy is shown in example (13) for Bilinarra (Pama-Nyungan, Australia), which expresses conditionals by the relativization of the protasis.

(10) *[wà-**á** -gám-yà]*$_{PRO}$ *[e-èké-gúnówó]*$_{APO}$
s:2SG-COND-greet-O:3SG s:3SG-NEG.FUT-answer
'If you greet X, X will not respond.'
Oko (Atoyebi 2010: 94)

(11) *[ajalwarra darrug arrunje-**ngay**]*$_{PRO}$ *[barramara]*$_{APO}$
rain falls it.does-**NMLZ** you.tell.me
'If it rains, tell me.'
Kwini (McGregor 1993: 55)

(12) *[páki má*^{N}*-rmɨ=**ka**]*$_{PRO}$*; [ini-t-r-í-t*^{j}*aram]*$_{APO}$.
peccary kill-2PL.SS=**TOP** bring-APPL-1SG.O-PFV-2PL.S:IMP
'If you kill a peccary, bring it to me.'
Shiwiar (Kohlberger 2020: 195)

(13) *[Nyila=ma=rna=nga warlagu=ma ba-rru guliyan=ma]*$_{APO}$
that=TOP=1MIN.S=DUB dog=TOP hit-POT dangerous=TOP
*[**nyamu**=yi=nga baya-wu]*$_{PRO}$.
REL=1MIN.O=DUB bite-POT
'I'll hit the agressive dog, if it bites me.'
Bilinarra (Meakins & Nordlinger 2013: 307)

For the analysis presented in Section 5, verbal conditional markers such as *-a-* in Oko shown in (10) count as dedicated conditional constructions, since their primary function is the expression of conditionality. The marking strategies shown in (11), (12) and (13), on the other hand, have other primary functions (i.e. nominalizing, topicalizing and relativizing an event, respectively). Therefore, they do not count

towards the number of dedicated conditional constructions for our analysis in Section 5.

Another common expression used especially for real and predictive conditionals is a temporal clause. According to Thompson, Longacre & Hwang (2007), this is often found in Austronesian languages and in the macro area of Papunesia in general. Also Martowicz. (2011: 278) shows that languages in that area, i.e. in New Guinea and Australia tend to show a lower degree of expliciteness for conditional expressions than in other areas of the world. To give an example, we can see that the temporal subordinator *xən* in Oksapmin (Nuclear Trans New Guinea, Papua New Guinea) can be used to express both a temporal (14) and a conditional context (15).

(14) *was* *n-x-ti-pel=**xən*** *nox* *skul* *xəm*
wash 1/2.O-make-PFV-FUT.PL=**SBRD** 1SG school down
əp-di-p
come-PFV-EVID.PST.SG
'After they washed me, I came down to school.'
Oksapmin (Loughnane 2009: 442)

(15) *[dit* *blel* *mox* *o=m-de-m* *s-ja=**xən**]*$_{PRO}$
1DU.INCL child ANAPH leave=PROX.O-make-SEQ go-PRS.PL=**SBRD**
[ixil *i=n-x-ti-pli=xən=o]*$_{APO}$
3PL angry=1/2.O-make-PFV-FUT.PL=IRR=QUOT
'If we leave the child behind and go, they might be angry with us.'
Oksapmin (Loughnane 2009: 433)

The functional extension of temporal markers or constructions to conditionality has also been noted from a grammaticalization perspective; durative or non-punctual temporal expressions are one of the most common sources for conditional markers identified in Traugott (1985). Conditional expressions that originate from temporal expressions such as the one shown in (15) do not count towards the number of dedicated conditional constructions either, their main function being the expression of temporal relations.

3 Socio-linguistic factors shaping language structure

3.1 Effects of population size and structure

There is ample evidence for the impact of various extra-linguistic factors on language structure and grammar. An early typological study of the role of socio-linguistic factors on grammatical structures was done by Perkins (1992), who found an effect of cultural properties on the systems of deictic expressions. Smaller, more intimate societies with stronger social ties between members were shown to have more complex deictic systems, as they rely on more knowledge shared between the members of the speech community. Larger societies were shown to have less complex deictic systems, which was explained in terms of loser ties between members and thus less shared knowledge between any two speakers of the community. This general observation that "societies of intimates" and "societies of strangers" develop languages with systematic differences due to differences in their social structures has been discussed in many other typologically-oriented studies (e.g. De Busser & LaPolla 2015; Sampson, Gil & Trudgill 2009; Trudgill 2011b; Wray & Grace 2007).

In addition, there is a substantial body of quantitative work that investigates the influence of social structures on grammar. Most studies, especially earlier ones, used population sizes as a proxy for the structures of the speech communities. The choice of using population sizes is probably a practical one; even though obtaining accurate numbers for the size of various speech communities comes with many difficulties as well, it is still one of the easiest variable related to social complexity to quantify at a large scale.

More recent studies, however, have started to move away from this overly simplistic representation and have tried to include information on especially L2-speaker proportions. While this may still not be sufficient to accurately capture social structures based on what we know from the theoretical literature, L2-speaker proportions seem to be an equally or even more relevant socio-linguistic predictor of language structure than population sizes. For instance, Bentz & Winter (2013) find that the L2-speaker proportion is a better predictor of the size of nominal case systems than the number of speakers. However, Sinnemäki & Di Garbo (2018) find that combining the information of population size and the proportion of L2-speakers leads to better predictions of verbal inflectional synthesis. The study by Sinnemäki (2020) is remarkable in that it analyzes the interaction of phylogenetic, areal, socio-linguistic (language-external) and language-internal factors (word order) in their influence on the development of complex case systems. Indeed, he finds complex interactions

between these different factors, which, together with other more recent findings, also serve as a motivation for the present study.

Other important case studies besides the ones mentioned above are Lupyan & Dale (2010) and Sinnemäki (2009), who have used typological datasets to show that larger population sizes tend to be associated with less complex inflectional morphology. Dale & Lupyan (2012) and Nettle (2012) find similar effects of population size based on computer simulations. There are also a number of smaller case studies focusing on selected languages or language families reporting similar tendencies. For instance, DeLancey (2014) suggests that socio-linguistic factors can account for the development of analytical vs. synthetic structures in different Tibeto-Burman languages. Kusters (2003), focusing on Arabic, Scandinavian, Quechua and Swahili, also shows that the number of L2-speakers, the social tightness of the speech community and the prestige of a language can shape the linguistic complexity of verbal inflection morphology.

Besides population sizes, Sinnemäki & Di Garbo (2018) show that L2-speaker proportions are an important additional predictor of verbal morphological complexity and grammatical gender. Looking at the number of nominal cases, Bentz & Winter (2013) find that the proportion of L2-speakers is a better socio-linguistic predictor compared to the population size; the higher the L2-proportion, the fewer case distinctions languages tend to have.[6]

Apart from work on structural complexity, previous studies have also found that vocabulary size is affected by socio-linguistic factors. Larger speech communities, which tend to have less complex structures, were shown to have larger vocabulary sizes than smaller speech communities (e.g. Reali, Chater & Christiansen 2018).

3.2 Modality effects: Written and spoken language

There is also a long tradition of investigating the impact of the modality, e.g. written vs. spoken language, on grammar. Modality effects are relevant on the synchronic as well as on the diachronic level. On the one hand, different modalities of the same language can show different preferences for certain linguistic structures. For instance, the preference against complex syntactic structures in spoken as opposed

6 Also phonological complexity in the form of phoneme inventory sizes has been suggested to correlate with social complexity. For instance, Trudgill (2004) argued that population size, the degree of linguistic contact, the tightness of social networks and the degree of social stability can influence the size of phoneme inventories. See Donohue & Nichols (e.g. 2011); Moran, McCloy & Wright (2012); Pericliev (2004); Wichmann, Rama & Holman (2011) for quantitative crosslinguistic studies of the association between phoneme inventory size and population size.

to written varieties of English was already shown by various authors early on (e.g. Halliday 1994; Miller & Weinert 1998; Pawley & Syder 1983; Redeker 1984; Tannen 1982). Especially the works by Biber, analyzing data from mainly English, but also Somali, Tuvaluan and Korean, revealed a more complex interaction between modalities and registers, leading to systematic structural differences between varieties of the same language (Biber 1995, 2006; 2009).[7]

The written modality can also shape language structure in its long-term availability or in the form of a literature tradition.[8] Here, the availability and use of clause-combining devices is especially relevant for conditional constructions. In his seminal work, Ong (1982) discusses the following structural properties that languages with a primarily oral tradition have: (i) additive rather than subordinative syntactic structures, (ii) aggregative expressions, i.e. the use of epiteths or parallel structures and (iii) redundancy and repetition in order to ensure that both the speaker and the hearer keep track with the discourse.[9] He thus already notes that the use of complex syntactic structures with dependent clauses is favoured by written uses of languages, whereas chained clauses with no syntactic dependencies and repetitions are a typical property of spoken language. Over the course of time, then, the use of such structures is conventionalised. This in turn is argued to lead to systematic syntactic differences between languages with an orality tradition vs. languages with a literature tradition.

Related to that, Mithun (1984) discusses corpus data from Guwinggu (Gunwinyguan, Australia), Mohawk (Iroquoian, Canada & USA) and Kathlamet (Chinookan, USA), showing that these languages use deictic markers and independent clauses for what is usually expressed by complex clauses (matrix clauses with relative, complement or adverbial clauses) in English. She argues that this difference in the use and availability of syntactic subordination can, at least in part, be accounted for by the development of a literature tradition in languages like English. This then leads to the observable pattern that languages with a literature tradition tend to make use of more subordination than languages with a primarily oral tradition.

Furthermore, Biber (2006), Chafe (1982), Mithun (1984) and Ong (1982) argue that these modality differences can lead to systematic typological differences across

7 See Dąbrowska (2020) for a recent overview of such synchronic, individual effects of writing on language.

8 We follow Ong (1982: 1–3) and use the notion of "literacy" and "literature tradition" to refer to a written tradition only, which is opposed to a culture of oral traditions, which we will refer to as "orality" or "oralility tradition".

9 Ong (1982) discusses many other typical properties of languages with oral traditions, which are however less relevant for the purposes of the present paper.

languages. From the speaker's perspective, written language can be planned ahead more carefully than spoken language, and written language can also be adjusted, which is not possible in the spoken modality. Spoken language is thus typically more spontaneous and less planned. From the addressee's perspective, reading is usually much faster than listening to spoken language. Requiring less time makes it cognitively easier for the reader to keep all the parts of a complex sentence in their working memory, which may be more difficult with slower, spoken language.

In addition, subordination could also be required in written texts, for which much less context is provided by the information contained in discourse context of spoken language. The link between utterances can often successfully be conveyed by the use of prosodic devices in spoken language, as both the speaker and the hearer share much more information from the discourse situation itself. This is not necessarily the case for written texts, which may need to compensate for the lack of context and be much more explicit in how certain ideas, expressed as different clauses, are related to each other. Furthermore, Deutscher (2000: 182) points out that writing allows for the expression of more complex concepts and can ultimately lead to more complex communicative patterns that can influence language structures independently of the modality of use over the course of time.

Similarly, in his analysis of adverbial constructions in European languages, Kortmann (1997) finds systematic correspondences between the most elaborate systems of adverbial subordinators and the literature tradition of the languages. Languages with fewer adverbial subordinators are also the languages with relatively young or no literature traditions, namely Romani, Talysh, Karaim, Sardinian, Manx, Gagauz, Ossetic, Udmurt, Komi, Nenets (Kortmann 1997: 254–255). He also points to the distance between the writer and the reader in written communication and the lack of extra-linguistic clues. Kortmann argues that those characteristics lead to a higher degree of syntactic explicitness being necessary for successful communication.

Another piece of evidence pointing in the same direction is that subordinators are sometimes borrowed from national languages with a writing tradition into other, local languages as a consequence of language contact. To give one example, Bakker & Hekking (2012) show how Otomi (Otomanguean, Mexico) borrowed a number of conjunctions and subjunctions from Spanish. Otomi is a predominantly oral language, having been in contact with Spanish for about 500 years. However, until recently, the Otomi communities could stay fairly monolingual due to their remote locations, and widespread bilingualism with Spanish in the Otomi communities only started around 1950. Bakker & Hekking (2012) show that the combination of clauses in Otomi can be left implicit in many cases and that existing conjunctions and subjunctions have rather broad semantic functions. Due to contact with Spanish, however, various explicit markers to combine clauses have been borrowed

from Spanish, and the existing Otomi markers have also become semantically more restrictive and specialised over time.

3.3 Implications for the number of conditional constructions

The effects that especially population size and L2-speaker proportions appear to have on the structural complexity of languages suggest that we may also find effects on the types of conditional constructions, which in turn could influence the number of dedicated constructions. Given that we see effects of population size on the morphosyntactic complexity of languages, we may expect languages with smaller speech communities to be more likely to make use of verbal inflection to encode conditionality, while languages with larger speech communities may tend to use syntactic markers. Once a system already has a morphological marker as a part of the verbal paradigm, its availability may in turn lead to fewer additional syntactic constructions. In languages with no morphological means to express conditionality, the development of syntactic conditional constructions may be favoured.

Similarly to the situation of Otomi mentioned in the previous section, we expect borrowing of conditional constructions and markers to take place in settings where most speakers of the community are bilingual. Only a general dominance of both languages in the community will allow for code-switching and language mixing, which is how a syntactic marker or construction could be borrowed from one language into another. We thus expect a higher number of conditional constructions in languages with smaller population sizes in those cases in which its speakers are multilingual and use another language with a more official status and a writing tradition.

Besides the effects of writing mentioned in Section 3.2, there is also evidence that the expression of conditionality, together with anteriority, is prone to be influenced by the written use of a language. In her 2011 study, Martowicz examined the properties of the expression of anteriority, causality, purpose and conditionality in a sample of 84 languages. She found an association between the grammaticalization, lexicalization and explicitness of conditional markers and various sociolinguistic factors: “By contrast, the evidence gathered for anteriority and conditionality suggest [sic.] that encoding of these two relations is very prone to the influence of socio-cultural factors” (Martowicz 2011: 310). Especially the level of written form development, the presence of radio and TV broadcasts, the number of speakers and the type of society (predominantly non-urban, mixed, predominantly urban) were found to be associated with the degree of explicitness of conditional constructions (Martowicz 2011: 312). This result also suggests that we should find a higher

number of dedicated (i.e. explicit) conditional constructions in languages that are used in writing and formal ways of communication, which is typical for languages with larger population sizes. At the same time, we may also expect that the absence of dedicated conditional constructions is more likely in oral languages of small communities which do not have a writing tradition.

4 Sample and annotation

Our dataset consists of 374 languages from 118 top-level families across the six macro-areas of Africa, Australia, Eurasia, North America, Papunesia and South America as used in Glottolog (Hammarström et al. 2021).[10] We included 50 languages for each macro-area.[11] Because of more data being available, including the results from Khrakovskij (2005), we have data from 127 languages for Eurasia (cf. Section 5.2 for how we control for a potential phylogenetic and contact bias). The relevant information was taken from reference grammars and language descriptions. If possible, both the grammatical and the socio-linguistic information was extracted from the language descriptions. For some languages, appropriate online databases were consulted for the sociological and demographic details.[12]

Figure 1 gives an overview of the distribution of the languages in the sample. The languages are coloured according to the number of dedicated conditional constructions they have, ranging from 0 constructions (dark) to 15 (light). The map already shows that most of the world's languages have a small number of constructions, higher numbers appear to be especially common in Europe and to a lesser extent in Asia and Africa. Due to the complexity of uses and variation in the descriptions, we did not exclude or distinguish conditional constructions according to their type (real, hypothetical and counterfactual). However, we only counted in

10 All data, sources and the code are provided in the online supplementary materials: https://gitlab.com/mguzmann89/conditionals-paper-lb-mgn-so.

11 The macro areas used in WALS and Glottolog are designed in a way that they are maximally independent of each other and comparable in terms of their genetic and typological diversity (Hammarström & Donohue 2014: 169).

12 For instance, we consulted the AustLang resource for Australian languages (https://collection.aiatsis.gov.au/austlang/search), the Endangered Languages Project (https://www.endangeredlanguages.com/), as well as census data, e.g. the "Report on the Status of B.C. First Nations Languages" (https://fpcc.ca/wp-content/uploads/2020/07/FPCC-LanguageReport-180716-WEB.pdf), the Mexican "Censo de Población y Vivienda 2020" (https://cuentame.inegi.org.mx/hipertexto/todas_lenguas.htm), and the "Philippine Statistics Authority 2014" (https://psa.gov.ph/sites/default/files/2014%20PIF_0.pdf).

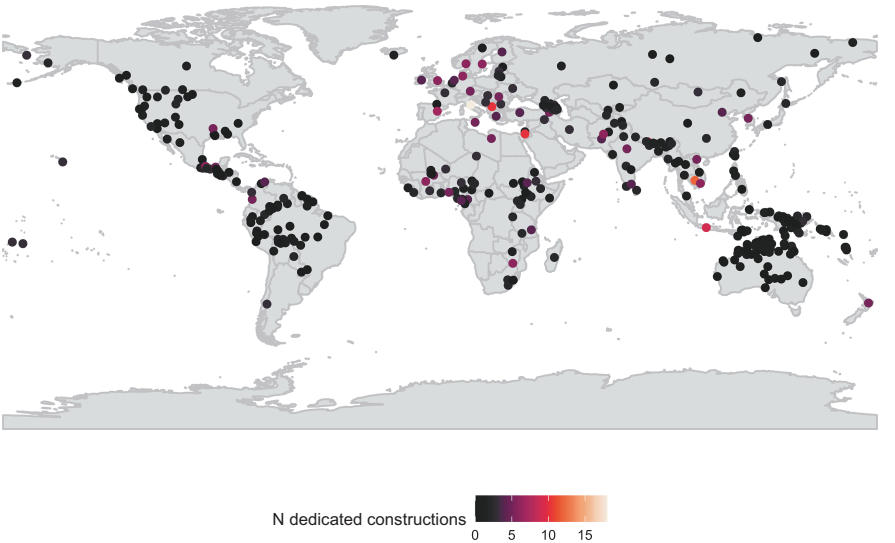

Figure 1: Map of the sample.

overtly marked, dedicated conditional constructions. In other words, constructions corresponding to juxtapposed clauses which are not formally marked (cf. example (9)) are not treated as a dedicated conditional construction. Neither are formally marked constructions counted in if they are used in other contexts, e.g. as a temporal or relative construction (cf. examples (11)-(15) from Section 2). This led to some languages of the dataset having 0 dedicated conditional constructions.

To give an example of how we counted dedicated conditional constructions, we can consider Bariai (Austronesian, Papua New Guinea). Gallagher & Baehr (2005: 161) discuss two conditional markers, *oangga* 'if/when' and *padam* 'only if'. Both markers occur in the protasis and they can additionally be accompanied by *eine* 'then' or *tota* 'therefore' in the apodosis. Example (16a) shows the use of *oannga* 'if/when' on its own, and (16b) together with *tota* 'therefore' in the apodosis. In (17), we see the use of *padam* in the protasis together with *eine* 'then' in the apodosis. Because the markers *eine* and *tota* are described as being optionally used in addition to one of the two conditional markers in the protasis, we treat cases such as (16a) and (16b) as two variants of a single construction. However, since *oannga* is used to express both temporal and conditional relations without necessarily disambiguating the two interpretations in a given context, we do not treat the construction containing *oannga* as a dedicated conditional construction. Thus, only the construction with *padam*, shown in (17) counts towards the number of dedicated constructions in Bariai.

(16) a. *[Ei ga i-pul ei mulian]$_{APO}$* ***[oannga*** *i-gera go.]$_{PRO}$*
3SG FUT s.3SG-turn 3SG back **if/when** s.3SG-see O.2SG
'He will turn back when he sees you.'

b. ***[Oangga*** *a-ean-ean anngа toaiua dadanga-i,]$_{PRO}$*
if/when s.1PL.EXCL-RED-eat food that garden-LOC
[oangga kus, tota amai anngа eta mao.]$_{APO}$
if/when be.done therefore POSS.1PL.EXCL food one.IRR not
'When we are eating that food in the garden, if it's gone, then we don't have any (more) food.'
Bariai (Gallagher & Baehr 2005: 161)

(17) ***[Padam*** *le-da eau i-eno-no,]$_{PRO}$ [eina*
if.only POSS-1PL.INCL water s.3SG-lay-RED then.there.2
ta-kona-ona.]$_{APO}$
s.1PL.INCL-hook-red
'If only some of our fuel was left, then we (could) hook-fish.'
Bariai (Gallagher & Baehr 2005: 161)

The socio-linguistic and extra-linguistic factors that we annotated for each language can be seen in Table 1.

Table 1: Socio-linguistic and extra-linguistic variables annotated.

variable	values
N speakers	(log) number of speakers
N L2-speakers	number of L2-speakers
multilinguals	no < some < many < most < all
literature	no literature, literature
writing	no < little < yes
education	no < language classes < little < yes
phylo	phylogenetic tree (taken from Glottolog)
latitude	latitude of the language's location (taken from Glottolog)
longitude	longitude of the language's location (taken from Glottolog)
biblio	grammar length measured in number of pages

Besides the number of speakers, there is evidence pointing towards the importance of L2-speaker proportions when examining the effects of socio-linguistic factors on grammar (cf. Section 3.1). However, similarly to previous studies, we had difficulties gathering sufficient information on the number of L2-speakers of all languages in the sample. We could only find reliable numbers for 43 out of 374 languages.

Some grammars describe the language as a lingua franca of the region, but they do not necessarily give any numbers of L2-speakers.[13]

Because exact numbers of L2-speakers were difficult to come by for most languages of the sample, we included a less exact measure of the proportion of multilinguals in the speech community, distinguishing between 5 ordinal values of no < some < many < most < all speakers being multilingual. We could annotate this information for most of the languages, as most grammars provide a rough estimate of the proportion of multilingual speakers. Only for 9 languages of the sample, we could not determine the proportion of multilingual speakers from the sources; we annotated their level as "unknown". We are aware that such an ordinal representation of the proportion of multilingual speakers does not correspond to the proportions of L2-speakers in the strict sense, but we included this variable for practical and exploratory reasons. Still, we hypothesize that a high degree of multilingualism could reflect a high degree of language contact, which could result in more constructions due to borrowing and calquing.

In addition to the information on speaker numbers, we annotated the following three socio-linguistic variables: the availability of a literature tradition, the use of the language in writing and in the educational system. Ideally, we would have used a much more fine-grained distinction, including for instance the presence of the language in TV, in radio, in newspapers, the use of the language in legal circumstances, etc., similarly to the variables used by Martowicz (2011). Unfortunately, including this information for a large crosslinguistic sample is hardly possible at the moment–this kind of information is only available for a few languages from the sample. For the purposes of the present study, we prioritized sample size over a rich and detailed socio-linguistic annotation as a first approach that can be supplemented by a smaller but more detailed follow-up study.

For the availability of a literature tradition, we simply made a binary distinction between the presence vs. the absence thereof. Whenever only a bible translation (or an equivalent translation of a religious text) was available, the language was annotated as having no literature tradition. Only if the community was described as producing written literature of their own accord, the language was annotated as having a literature tradition.

In addition to the availability of a literature tradition, we also annotated the extent to which a language is currently used in writing. Most language descriptions note in detail whether an orthography exists and whether it is used productively by the community. If the language had an alphabet but the community scarcely

13 This is not to criticize the authors of the grammar; rather, it shows how difficult it is to quantify the number of L2-speakers even for experts on a given language.

used it, we marked its use in writing as 'little'. The same holds if there were only translated texts such as the bible. If the orthography was solely used for scientific purposes or no orthography existed, the language was annotated as not being used in writing. Only if the orthography was accepted and used by the community to produce a variety of written texts, the language was considered to be fully used in writing.

For the use in education, we distinguished between four values. If the language was used as the medium of instructions in schooling and/or in higher education, we annotated it as used in education. If the use of the language as the medium of instruction only had a very limited range, e.g. only in the first classes of elementary school, its use in education was annotated as 'little'. If only language classes (for children and/or adults) but no other education in the language was available, we annotated it as having language classes. If there was neither formal instruction in a language nor language classes, we marked it as not being used in education.

As mentioned in Section 3.3, we hypothesize that both the availability of a literature tradition and its use in writing makes a higher number of conditional constructions more likely. The use of the language in the educational system is very likely correlated with the other two variables; we included it because we did not know a priori which variable proved to be the most informative predictor (and because we had sufficient information about this variable for the languages in the sample). As was mentioned in Section 3.2, we know that this modality effect holds for adverbial markers in general; we can assume that it is the written modality that requires conditionals to be made more explicit as opposed to the spoken modality, where conditionals can be left morpho-syntactically unmarked and where the discourse context and prosody play a more important role.

We know that the degree to which languages are fully described varies drastically across areas and families and individual languages, and it is very likely that we miss linguistic details on conditional constructions in a given language simply because there is only a single description which has to focus on many different aspects of the language. In order to account for such a potential bibliographical bias (cf. Bakker 2010), we also annotated the length of the grammars used. The length was measured in number of pages.[14] In case more than one source per language was consulted, we used the longest description.

Finally, we included two other extra-linguistic variables from Glottolog, namely the phylogenetic information and the coordinates of the languages. We used these

14 Alternatively, one could have coded the length of the sections or chapters on conditional constructions. While more precise in theory, we did not opt for this solution because conditional constructions were often treated in more than one section in the descriptions.

two variables to account for phylogenetic and contact biases in our model, which will be explained in more detail in Section 5.2.

5 Results

5.1 Overall distributions

In this section, we will give an overview of the raw distributions, showing the relevant patterns to examine the association between the number of conditional constructions and various socio-linguistic factors. Figure 2 shows how the number of conditional constructions is associated with the log number of speakers (left) and the degree of multilingualism (right).

As we can see in the left plot of Figure 2, the number of speakers appears to be weakly associated with the number of conditional constructions in that all languages in the dataset with a high number of constructions (>10) also have larger population sizes. Indeed, the two measures have a moderate positive correlation of 0.46. However, fewer conditional constructions are found independently of the number of speakers.

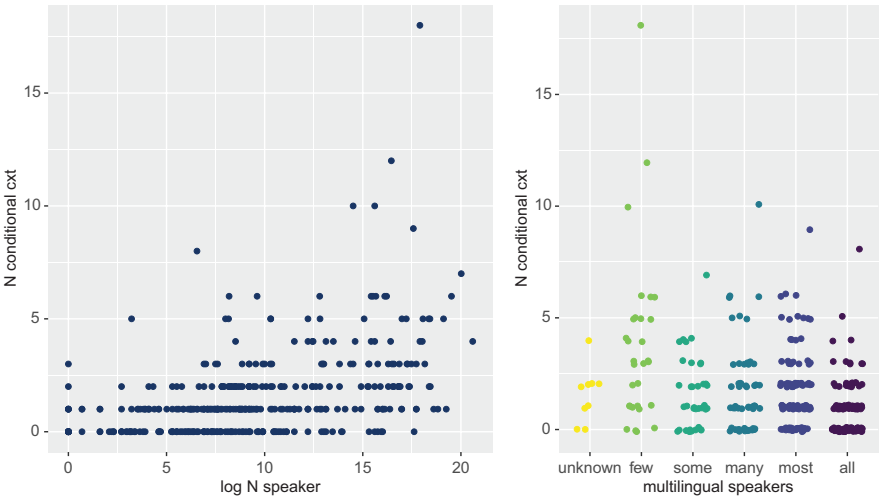

Figure 2: The number of conditional constructions by log N speaker (left) & multilinguals (right).

Regarding the degree of multilingualism or rather the importance and use of other languages, the right plot in Figure 2 does not show any clear trends. If at all, it appears that languages with fewer multilingual speakers may have slightly more conditional constructions. This may be due to the fact that the languages with few multilingual speakers usually correspond to national languages with larger population sizes, and the languages with many multilingual speakers are often those that have only a small number of speakers.

Figure 3 shows the associations between the number of conditional constructions and three socio-linguistic variables: the use of the language in education (left), in writing (center) and the availability of a literature tradition (right). For all three of those variables, we see a very weak association with the number of conditional constructions. Again, the association rather concerns high numbers of constructions; they only occur in those languages that are fully used in the educational system and in writing and that have a literature tradition. At the same time, lower numbers of conditional constructions are found across all categories of the three socio-linguistic variables.

One of the issues with the associations seen in Figures 2 and 3 above is that the socio-linguistic values that we want to examine as predictors of linguistic properties are correlated with each other. To show a few examples, Figure 4 plots `log N speakers` against `education`, `writing` and `literature`. We see a very clear association with the number of speakers of a language; the more speakers a language has, the higher its tendency to be used in education and writing and to have a literature tradition.

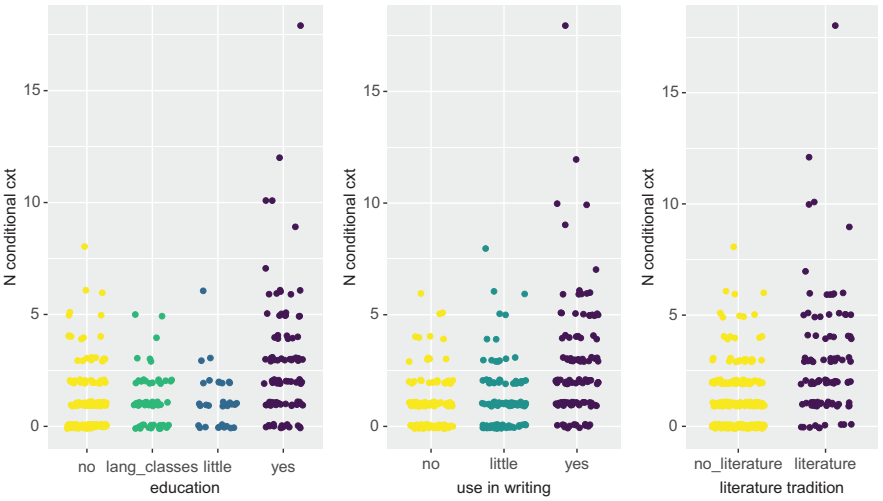

Figure 3: The number of conditional constructions by education (left) & writing (center) & literature (right).

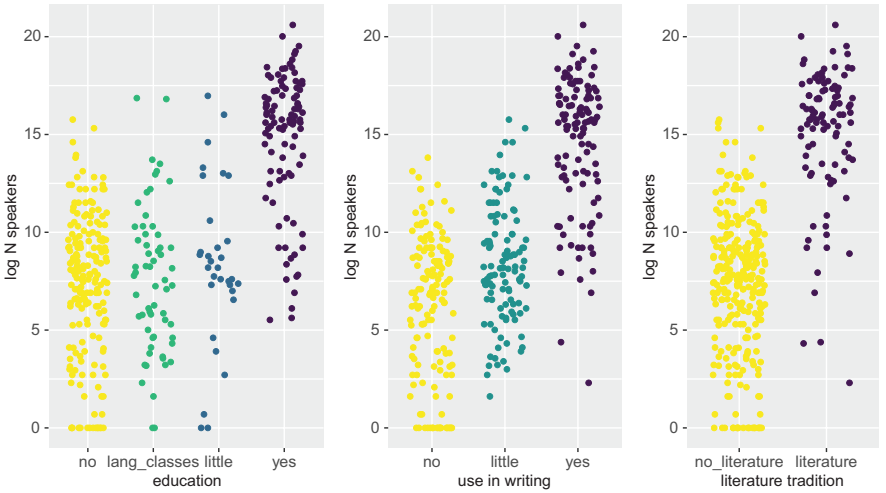

Figure 4: Log N speakers by education (left) & writing (center) & literature (right).

Another bias to consider is the bibliographical bias. To ensure that we do not find a higher number of constructions simply because of an overall more thorough language description, we checked for the association between the number of conditional constructions and the length of grammars measured in number of pages. The association between the two variables can be seen in Figure 5.

The distribution in Figure 5 shows that there does not seem to be any clear correlation between those two variables. Indeed, their correlation is 0.005, which points to virtually no association between the two variables.

5.2 Modelling the number of conditional constructions

The aim of this section is to analyze the effect of various socio-linguistic variables on the number of conditional constructions, taking into account phylogenetic, contact and bibliographical effects. As was mentioned in the previous section, we used grammar length as an approximation of the bibliographical control. For phylogenetic and contact controls, we followed the method described in Guzmán Naranjo & Becker (2021). In order to control for contact effects, we used a two-dimensional Gaussian Process with the coordinates of the languages. The basic idea behind the Gaussian Process term is that languages that are spoken in closer proximity are more likely to influence each other than languages that are spoken with larger distances between them. While this still is a very crude approximation, the Gaussian Process has the advantage that we do not have to assume a constant effect of dis-

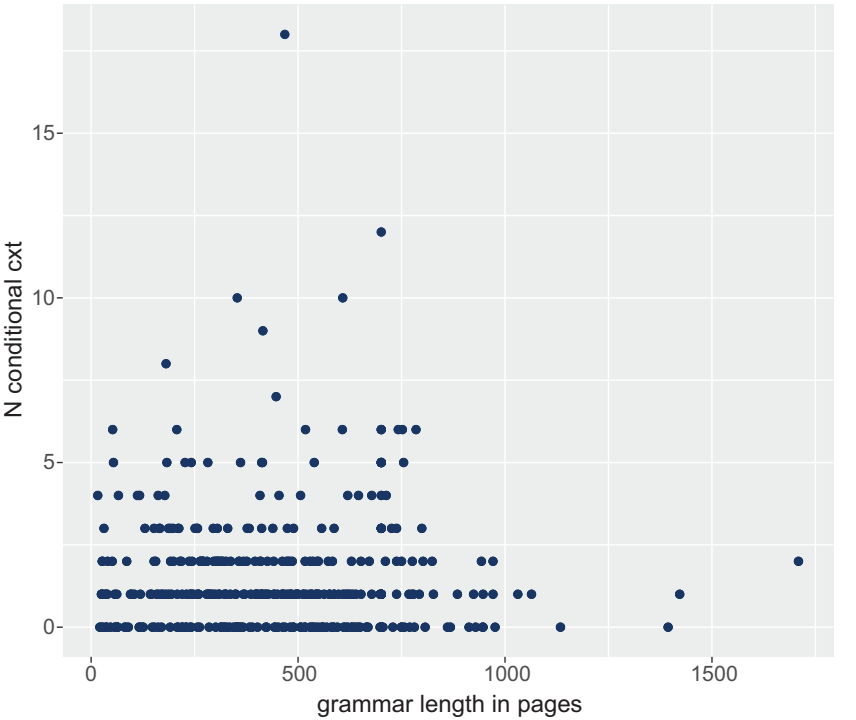

Figure 5: The number of conditional constructions by grammar length.

tance between two languages. In other words, based on the data itself, the model is able to find the distances relevant for contact between languages in a non-linear way.[15] The other additional control consists of a phylogenetic regression term, using the information of the entire phylogenetic trees of the languages to model phylogenetic effects. This method allows us to represent phylogenetic relations in a gradient way instead of grouping languages together at an arbitrarily chosen family level. The model thus forces the estimates of closely related languages to be more similar than those of less closely related languages.

As was shown in the preceding section, our socio-linguistic variables (`N speakers, education, writing, literature, multilinguals`) are heavily correlated. Because of that, including all relevant variables as predictors can lead to biased estimates in a model, and a more complex procedure of model selection is necessary.

15 This is important, since we know that the distance between two languages in contact depends on e.g. the linguistic density of the area; it differs substantially across different areas of the world.

Since we are dealing with count data, we used a Poisson distribution. We fitted two types of regression models with the controls described above.[16] The first model ("controls+5" model) includes all five socio-linguistic predictors as well as the phylogenetic, contact and bibliographical controls. The second type of models each includes one single predictor in addition to the three controls ("controls+1" models). By fitting these two types of models, we can compare the effects of the predictors in the presence of the other predictors (controls+5 model) to their effects in isolation (controls+1 models). This ensures that we do not miss an effect that the predictors could have in the presence of the other predictors and the controls.

Figures 6 to 10 show the conditional effects for the five predictors. Conditional effects correspond to the estimated effects drawn from the model predictions.[17] The red dots or lines represent the mean values of the posterior distribution of the number of conditional constructions, and the error bars or bands show the 95% uncertainty intervals. The uncertainty intervals correspond to the intervals that 95% of the posterior distribution falls into. This means that given the data and the model, we can be 95% certain that the number of conditional markers will fall in that interval. In each of the figures, the left plot shows the conditional effects of the predictor in the controls+5 model, i.e. the one including all five predictors. The right plots all show the conditional effects of the predictor in the controls+1 model, where no additional socio-linguistic predictor is used. As expected, across all five predictors, the controls+1 model with only one of the predictors (right) shows

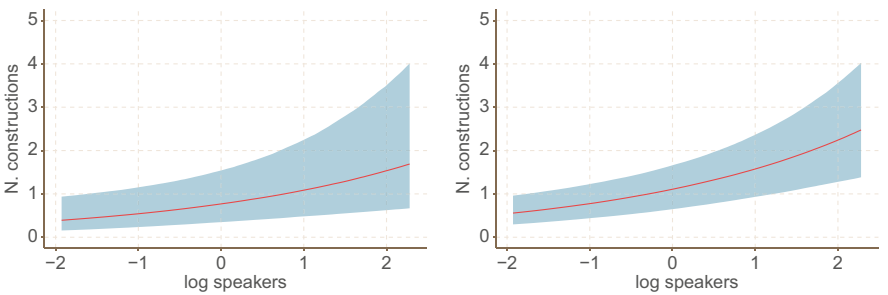

Figure 6: Conditional effects for N_speaker: controls+5 model (left) and controls+1 model (right).

16 The models were fitted using Bayesian methods with Stan (Carpenter et al. 2017) and the *brms package* Bürkner 2017) in R (R Core Team 2021). See the supplementary materials for the code.

17 Note that we centered and scaled both the log value of the number of speakers and the grammar length, which is why there are negative values for number of speakers in the conditional effects plot.

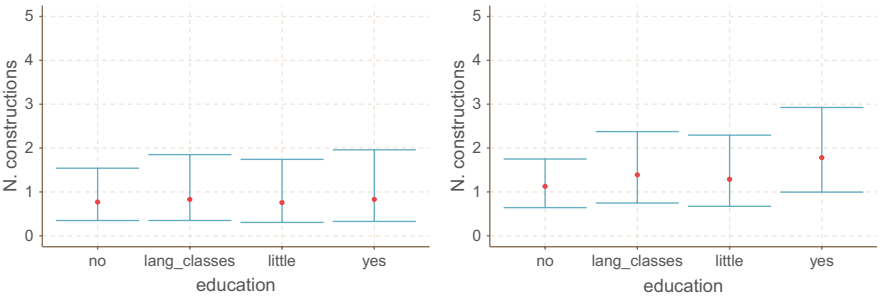

Figure 7: Conditional effects for education: controls+5 model (left) and controls+1 model (right).

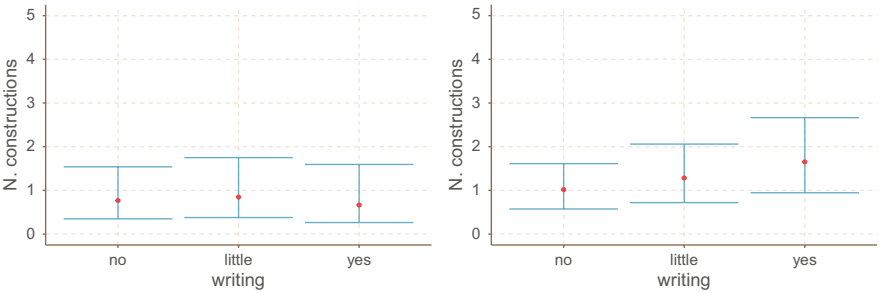

Figure 8: Conditional effects for writing: controls+5 model (left) and controls+1 model (right).

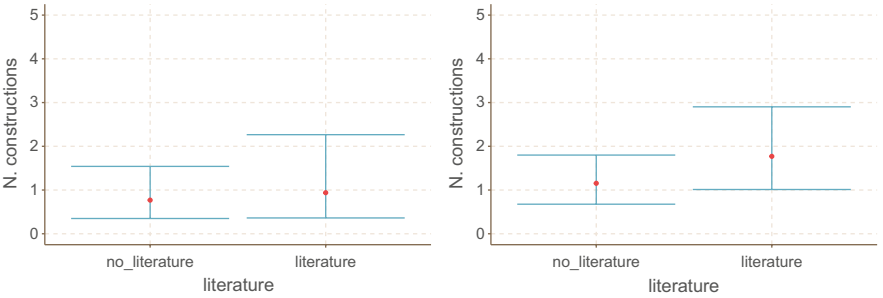

Figure 9: Conditional effects for literature: controls+5 model (left) and controls+1 model (right).

larger effect sizes than the controls+5 model (left) which includes all five socio-linguistic predictors. The important point here is that even though effect sizes are larger, the uncertainty intervals are so large that the models suggest a non-effect. The only predictor which has a weak effect in the presence of the controls is the number of speakers shown in Figure 6. For the other predictors, the uncertainty intervals strongly suggest that there is in fact no effect. We can thus say that, based

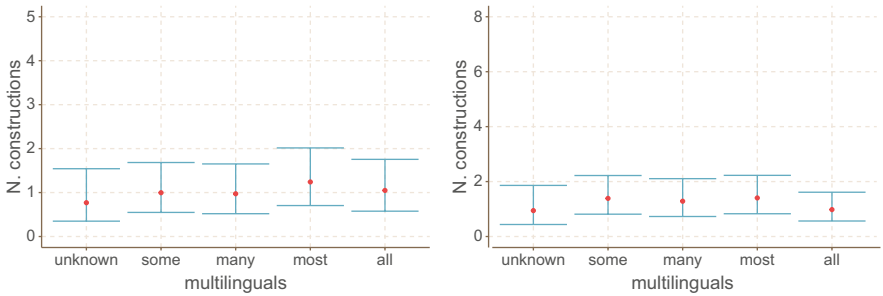

Figure 10: Conditional effects for multilinguals: controls+5 model (left) and controls+1 model (right).

on the conditional effects of the models discussed so far, only the number of speakers appears to have a weak influence on the number of conditional constructions if phylogenetic, contact and bibliographical biases are controlled for. All other socio-linguistic predictors, i.e. the use in education, in writing, the availability of a literature tradition and the proportion of multilingual speakers do not appear to be informative for the number of conditional constructions.

In addition to analyzing the effect sizes of the predictors, we can also assess and compare the predictive power of the models in order to select a final model that is the best one to generalize to new data from our sample. To do so, we compared model performance using 10-fold cross-validation following Vehtari, Gelman & Gabry (2017). 10-fold cross-validation re-fits a model leaving out 10% of the data at a time and then predicts those data points. This allows to evaluate the overall model performance against data which has not been used to train the model. To compare models we use the ELPD value (theoretical expected log pointwise predictive density), which measures how well a model is expected to predict a new dataset. The absolute value itself is not relevant here; it is rather the relative difference of ELPD values that can be used to compare models in terms of their predictive power.

Table 2 shows the comparison of a number of models with different combinations of predictors. The model with neither predictors nor controls (10) serves as a no-information-baseline, showing the predictive power of a model that does not have any information other than the overall distribution of the number of conditional constructions to predict from. The 10 models are ranked from highest (top, 1) to lowest (bottom, 10) predictive power. They are listed with their relative ELPD difference to the best performing model and with the standard error for that differ-

ence. The standard error helps to assess the ELPD difference between two models; only a difference larger than twice its standard error is likely to be meaningful.[18]

Table 2: Model comparison (10-fold cross-validation).

		ELPD difference	SE difference
1	phylo + contact + biblio + log N speaker	0.0	0.0
2	phylo + contact + biblio + literature	–4.2	4.1
3	phylo + contact + biblio + education	–4.3	3.9
4	contact	–7.7	8.5
5	phylo + contact + biblio	–9.8	5.3
6	all predictors	–12.7	3.5
7	phylo + contact + biblio + multilinguals	–17.7	4.5
8	phylo + contact + biblio + writing	–18.2	5.6
9	phylo	–53.7	8.8
10	no controls / predictors	–121.5	29.6

The model with the best performance in terms of predictive power is the model with all three controls and only `log N speakers` as a predictor (1). Adding any other predictor (and combinations thereof) does not seem to improve the model. While adding `literature` (2) or `education` (3) leads to better models than the one with controls only (5), the difference is too small for us to conclude that it is not due to random variation. Adding `multilinguals` (7) or `writing` (10) results in worse models in terms of performance. As for the socio-linguistic predictors, we can thus conclude that the best predictor is the number of speakers. It is also likely that `education` and `literature` only improve the model by virtue of being correlated proxies of the number of speakers.

An interesting point is that the model including the three controls (5) has almost as much predictive power as the model including only contact (4). This strongly points towards the possibility that the main driving force behind the distribution of our data is contact, more so than genetic effects or socio-linguistic factors.

18 A more conservative estimate is that the ELPD difference should be at least four times larger than its standard error because standard error estimates can be biased (cf. Vehtari, Gelman & Gabry 2017).

6 Discussion

As for the relation between the socio-linguistic variables, we showed that the number of speakers, the use of the language in education and in writing as well as the availability of a literature tradition are highly correlated with each other. This made it difficult to determine which variable was the best predictor of the number of conditional constructions in a given language. Our model comparisons in Section 5.2 showed that in our case, the most informative socio-linguistic predictor was the number of speakers. The best-performing model predicted more conditional constructions for languages with higher numbers of speakers, and less constructions for languages with smaller population sizes. Although the effect of the number of speakers was rather weak, the results of this study are somewhat in line with the results from earlier studies, reporting on associations between socio-linguistic factors and the availability of (a high number of) explicit adverbial subordinators (Kortmann 1997: 254–255) and the degree of lexicalization, grammaticalization and explicitness of conditionals (Martowicz 2011: 310). Yet, as the effect was shown to be very weak, we should probably be careful in assuming a strong and direct effect of socio-linguistics factors such as the written use of a language on complex syntax. There may be many single cases to evidence such effects, and we may find a crosslinguistic association in larger samples. Including controls for other biases, however, showed that the effects of socio-linguistic factors on complex (morpho-) syntax such as conditional constructions are probably much less direct and less strong than previously assumed.

One potential issue that is very hard to resolve in practice concerns the time depth of our information. All our variables represent the current situation of the language, and we would have to assume that we can generalize from that to the point in time in which the conditional constructions formed. It is possible that some of the uses of the language (e.g. in the educational system) are too recent in order to lead to changes in the language, and numbers of speakers may have drastically varied for other points in time. Related to that, we do not necessarily know when the constructions in question developed; conditional constructions (and probably also other complex constructions) may also have developed at different points in time in a given language. These issues likely also contribute to the rather weak association that we found, as it is not feasible to obtain accurate historic information on these variables for such a large dataset. A future small-scale study with a subset of the languages of the current sample where more historical data is available may be useful to verify whether or not we can work with such approximations.

Also, finding the strongest effects for number of speakers may be somewhat surprising, given the trend in quantitative socio-typology to move away from population sizes and to focus on the proportions of L2-speakers (Bentz & Winter 2013;

Sinnemäki 2020; Sinnemäki & Di Garbo 2018). This was one of our motivations to include the additional socio-linguistic variables in the first place. Although our data suggests that the number of speakers has a stronger effect on the number of conditional constructions in a language compared to the other variables, we cannot fully discard the latter. Since it was difficult to gather accurate and detailed information on the use of the language and its domains, it could also be the case that our additional socio-linguistic variables proved to be less informative due to the lack of consistent crosslinguistic documentation. It could well be that with more detailed data being available, variables such as the use in writing and the availability of a literature tradition could be included in a more fine-grained manner, which could in turn make it a more important predictor of the linguistic property analyzed. As we used those predictors in the present study, only the number of speakers was a numeric variable, whereas the other four socio-linguistic variables were ordinal or binary, i.e. of a different data type with inherently less information.

The danger of reducing complex socio-linguistic realities to population sizes was also expressed by Trudgill (2011a: 156) in the debate about a relation between population sizes and phoneme inventory sizes:

> My suggestion was very much that the five social factors could be expected, in combination, to have various kinds of influence on phoneme inventory size; it will never, I suggest, be sufficient to look at population figures alone. It is of course not surprising that Pericliev and other statistically-minded linguists have neglected this point and focussed on population size to the exclusion of the other factors, because the other factors are much less readily susceptible to quantification than community population size. But from my perspective this is actually a mistaken exercise – I see no socio-linguistic reason to suppose that population size alone will have any straightforward consequences for phoneme inventory size.

In the present study, we attempted to do justice to this very relevant objection. However, as long as the information on other socio-linguistic variables is not systematically included in language descriptions or linguistic databases such as Glottolog, it may well be that the number of speakers remains the most reliable socio-linguistic variable available on a large scale. Also, including information on the number of speakers is not necessarily a bad practice. In the light of more and more studies that uncover effects of socio-linguistic variables on grammatical structures, quantitative typological studies should include (at the very least) the number of speakers as a control, similarly to how macro areas or language families are generally controlled for in the modelling of typological data.

Besides analyzing the effect of socio-linguistic variables on the number of conditional constructions, we also included statistical controls for potential phylogenetic, contact and bibliographical biases. Our results showed that those control variables are equally important as the socio-linguistic ones. It is also important to

note that the contact control seemed to lead to a higher predictive power than the phylogenetic control. In other words, only controlling for language families is not enough –large-scale typological studies should always control for potential contact biases as well. We also saw in Section 5.2 that the effects of the socio-linguistic predictors were stronger in the absence of the controls. This means that part of the information of the socio-linguistic variables is already included in the controls; therefore, it is crucial for socio-typological studies to properly control for such additional effects. Finally, we also included a bibliographical bias control in our models. As far as we are aware, this is not yet a standard in quantitative typology, even though it is very easy to implement. Including such information in future studies can help to better understand the complex interactions of the different extra-linguistic variables.

7 Conclusion

In this paper, we analyzed the effects of several socio-linguistic variables on the number of conditional constructions while controlling for phylogenetic, contact and bibliographical biases. In order to allow for a more variegated picture than "simply" using population size, we included the following additional variables: the use of the language in education and in writing, and the availability of a literature tradition. Our objective was to be able to better explore the effect of written language and a tradition of literature on complex expressions such as conditionals. We chose conditionals as a testing ground because results from previous studies pointed towards an association between the use and availability of written language and number of explicit conditional constructions. Our results, however, suggested only a very weak association with the number of speakers when controlling for phylogenetic, contact and bibliographical biases. We showed that it is important to properly control for phylogenetic and contact relations, as they are at least as important (if not more important) predictors to determine the number of conditional constructions in a given language. We did not find any strong bibliographical bias in our data, but we believe that it is an important and easy-to-implement variable in any quantitative typological study. The other socio-linguistic predictors did not show any effects. We argued that this does not necessarily reflect the inherent nature of the associations; instead, it could well be due to the lack of detailed and consistent information on various socio-linguistic factors on a typological scale. Therefore, closing this information gap should be one of the priorities of of language description and typology.

Glosses

1	first person
2	second person
3	third person
ANAPH	anaphoric
APPL	applicative
COND	conditional
DAT	dative
DU	dual
DUB	dubitative
EVID	evidential
EXCL	exclusive
FUT	future
GEN	genitive
HAB	habitual
HON	honorific
IMP	imperative
INCL	inclusive
IRR	irrealis
LOC	locative
M	masculine
MIN	minimal
NEG	negative
NMLZ	nominalizer
O	object
PFV	perfective
PL	plural
POSS	possessive
POT	potential
PROX	proximate
PRS	present
PST	past
PTCP	participle
QUOT	quotative
REL	relativizer
RED	reduplication
S	subject
SBRD	subordinator
SEQ	sequential
SG	singular
SS	same subject
TOP	topic

References

Athanasiadou, Angeliki & René Dirven (eds.). 1997. *On conditionals again*. Amsterdam: Benjamins.

Atoyebi, Joseph Dele. 2010. *A reference grammar of Oko*. Köln: Köppe.

Bakker, Dik. 2010. Language sampling. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 100–127. Oxford: Oxford University Press.

Bakker, Dik & Ewald Hekking. 2012. Clause combining in Otomi before and after contact with Spanish. *Linguistic Discovery* 10(1). 42–61.

Bentz, Christian & Bodo Winter. 2013. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3. 1–27.

Biber, Douglas. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press.

Biber, Douglas. 2006. *University language: a corpus-based study of spoken and written registers*. Amsterdam: Benjamins.

Biber, Douglas. 2009. Are there linguistic consequences of literacy? comparing the potentials of language use in speech and writing. In *The Cambridge Handbook of Literacy*. Cambridge: Cambridge University Press.

Bürkner, Paul-Christian. 2017. Brms: an r package for bayesian multilevel models using stan. *Journal of Statistical Software* 80(1). 1–28.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1). 1–32.

Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen (ed.), *Spoken and written language: Exploring orality and literacy*, 35–53. Norwood, NJ: Ablex.

Comrie, Bernard. 1986. Conditionals: A typology. In Elizabeth Closs Traugott, Alice Ter Meulen, Judy Snitzer Reilly & Charles A. Ferguson (eds.), *On conditionals*, 77–99. Cambridge: Cambridge University Press.

Dąbrowska, Ewa. 2020. How writing changes language. In Anna Mauranen & Svetlana Vetchinnikova (eds.), *Language Change*: *The Impact of English as a Lingua Franca*, 75–94. Cambridge: Cambridge University Press.

Dale, Rick & Gary Lupyan. 2012. Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems* 15(3/4). 1150017.

De Busser, Rik & Randy LaPolla (eds.). 2015. *Language structure and environment: Social, cultural, and natural factors*. Amsterdam: Benjamins.

DeLancey, Scott. 2014. Sociolinguistic typology in North East India: A tale of two branches. *Journal of South Asian Languages and Linguistics* 1(1). 59–82.

Deutscher, Guy. 2000. *Syntactic change in Akkadian: The evolution of sentential complementation.* Oxford: Oxford University Press.

Diessel, Holger & Volker Gast. 2012. *Clause linkage in cross-linguistic perspective: data-driven approaches to cross-clausal syntax*. Berlin: De Gruyter Mouton.

Donohue, Mark & Johanna Nichols. 2011. Does phoneme inventory size correlate with population size? *Linguistic Typology* 15(2). 161–170.

Gallagher, Steve & Peirce Baehr. 2005. *Bariai grammar sketch*. Ukarumpa, Papua New Guinea: SIL.

Greenberg, Joseph. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg (ed.), *Universals of language*, 73–113. Cambridge, MA: MIT Press.

Guzmán Naranjo, Matías & Laura Becker. 2021. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670.
Haiman, John. 1978. Conditionals are topics. *Language* 54(3). 564–589.
Halliday, Michael. 1994. *Spoken and written language*. Oxford: Oxford University Press.
Hammarström, Harald & Mark Donohue. 2014. Some principles on the use of macro-areas in typological comparison. *Language Dynamics and Change* 4(1). 167–187.
Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2021. *Glottolog 4.4*. Leipzig: Max Planck Institute for the Science of Human History.
Hellwig, Birgit. 2011. *A grammar of Goemai.* Berlin: De Gruyter Mouton.
Hetterle, Katja. 2015. *Adverbial clauses in cross-linguistic perspective*. Berlin: De Gruyter Mouton.
Karlsson, Fred, Matti Miestamo & Kaius Sinnemäki. 2008. *Linguistic complexity. Typology, contact, change*. Amsterdam: Benjamins.
Khrakovskij, Viktor S. (ed.). 2005. *Typology of conditional constructions*. München: Lincom Europa.
Kirton, Jean F. & Bella Charlie. 1996. *Further aspects of the grammar of Yanyuwa, Northern Australia*. Canberra: Pacific Linguistics.
Kohlberger, Martin. 2020. *A grammatical description of Shiwiar*. Amsterdam: LOT.
Kortmann, Bernd (ed.). 1997. *Adverbial subordination: A typology and history of adverbial subordinators based on European languages.* Berlin: De Gruyter Mouton.
Kusters, Wouter. 2003. *Linguistic complexity: The influence of social change on verbal inflection.* Utrecht: LOT.
Ladd, D Robert, Seán G Roberts & Dan Dediu. 2015. Correlational studies in typological and historical linguistics. *Annual Review of Linguistics* 1. 221–241.
Loughnane, Robyn. 2009. *A grammar of Oksapmin*. Melbourne: University of Melbourne dissertation.
Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS ONE* 5(1). e8559.
Lupyan, Gary & Rick Dale. 2016. Why are there different languages? The role of adaptation in linguistic diversity. *Trends in Cognitive Sciences* 20(9). 649–660.
Martowicz, Anna. 2011. *The origin and functioning of circumstantial clause linkers: A crosslinguistic study*. Edinburgh: University of Edinburgh dissertation.
McGregor, William. 1993. *Gunin / Kwini*. München: Lincom Europa.
Meakins, Felicity & Rachel Nordlinger. 2013. *A grammar of Bilinarra, an Australian Aboriginal language of the Northern Territory*. Berlin: De Gruyter Mouton.
Miller, Jim & Regina Weinert. 1998. *Spontaneous spoken language: syntax and discourse.* Oxford: Clarendon.
Mithun, Marianne. 1984. How to avoid subordination. *Annual Meeting of the Berkeley Linguistics Society* 10(0). 493–509.
Moran, Steven, Daniel McCloy & Richard Wright. 2012. Revisiting population size vs. phoneme inventory size. *Language* 88(4). 877–893.
Nettle, Daniel. 2012. Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1597). 1829–1836.
Ong, Walter. 1982. *Orality and literacy: The technologizing of the word*. New York: Methuen.
Pawley, Andrew & Frances Hodgetts Syder. 1983. Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics* 7(5). 551–579.
Pericliev, Vladimir. 2004. There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language. *Linguistic Typology* 8(3). 376–383.
Perkins, Revere. 1992. *Deixis, grammar and culture*. Amsterdam: Benjamins.

Podlesskaya, Vera. 2001. Conditional constructions. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals. An international handbook*, vol. 2, 998–1010. Berlin: De Gruyter Mouton.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Manual. R Foundation for Statistical Computing. Vienna, Austria.

Reali, Florencia, Nick Chater & Morten Christiansen. 2018. Simpler grammar, larger vocabulary: How population size affects language. *Proceedings of the Royal Society B: Biological Sciences* 285(1871). 20172586.

Redeker, Gisela. 1984. On differences between spoken and written language. *Discourse Processes* 7. 43–55.

Sampson, Geoffrey, David Gil & Peter Trudgill (eds.). 2009. *Language complexity as an evolving variable*. Oxford: Oxford University Press.

Sinnemäki, Kaius. 2009. Complexity in core argument marking and population size. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 125–140. Oxford: Oxford University Press.

Sinnemäki, Kaius. 2020. Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach. *Journal of Historical Sociolinguistics* 6(2). 20191010.

Sinnemäki, Kaius & Francesca Di Garbo. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology* 9. 1141.

Tannen, Deborah. 1982. Oral and literate strategies in spoken and written narratives. *Language* 58(1). 1–21.

Thompson, Hanne-Ruth. 2012. *Bengali*. Amsterdam: Benjamins.

Thompson, Sandra A., Robert E. Longacre & Shin Ja J. Hwang. 2007. Adverbial clauses. In Timothy Shopen (ed.), *Language typology and syntactic description*, vol. 2, 237–300. Cambridge: Cambridge University Press.

Traugott, Elizabeth Closs. 1985. Conditional markers. In John Haiman (ed.), *Iconicity in syntax*, 289–307. Amsterdam: Benjamins.

Traugott, Elizabeth Closs, Alice Ter Meulen, Judy Snitzer Reilly & Charles A. Ferguson. 1986. *On conditionals*. Cambridge: Cambridge University Press.

Trudgill, Peter. 2004. Linguistic and social typology:The Austronesian migrations and phoneme inventories. *Linguistic Typology* 8(3). 305–320.

Trudgill, Peter. 2008. Linguistic and social typology. In J. K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The handbook of language variation and change*, 707–728. London: Wiley.

Trudgill, Peter. 2010. Contact and sociolinguistic typology. In Raymond Hickey (ed.), *The handbook of language contact*, 299–319. New York: Wiley Online Library.

Trudgill, Peter. 2011a. Social structure and phoneme inventories. *Linguistic Typology* 15(2). 155–160.

Trudgill, Peter. 2011b. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.

Vehtari, Aki, Andrew Gelman & Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5). 1413–1432.

Wichmann, Søren, Taraka Rama & Eric W. Holman. 2011. Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15(2). 177– 197.

Wierzbicka, Anna. 1996. *Semantics: Primes and universals*. Oxford: Oxford University Press.

Wray, Alison & George W. Grace. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117(3). 543–578.