Matías Guzmán Naranjo* and Laura Becker

Statistical bias control in typology

https://doi.org/10.1515/lingty-2021-0002
Received January 9, 2021; accepted October 5, 2021; published online November 17, 2021

Abstract: In this paper, we propose two new statistical controls for genealogical and areal bias in typological samples. Our test case being the effect of VO-order effect on affix position (prefixation vs. suffixation), we show how statistical modeling including a phylogenetic regression term (phylogenetic control) and a two-dimensional Gaussian Process (areal control) can be used to capture genealogical and areal effects in a large but unbalanced sample. We find that, once these biases are controlled for, VO-order has no effect on affix position. Another important finding, which is in line with previous studies, is that areal effects are as important as genealogical effects, emphasizing the importance of areal or contact control in typological studies built on language samples. On the other hand, we also show that strict probability sampling is not required with the statistical controls that we propose, as long as the sample is a variety sample large enough to cover different areas and families. This has the crucial practical consequence that it allows us to include as much of the available information as possible, without the need to artificially restrict the sample and potentially lose otherwise available information.

Keywords: affixation; bias control; phylogenetic regression; quantitative typology; sampling; word-order

1 Introduction

A common assumption in language typology is that systematic sampling is an essential part of controlling for genealogical and areal biases.[1] Statistical claims require independent observations, but since languages are related to each other both genealogically and through contact, we try to choose a (small) set of

1 This project was partially supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement 834050).

***Corresponding author: Matías Guzmán Naranjo /matias gusman naranxo/,** Eberhard Karls Universität Tübingen, Tübingen, Germany, E-mail: mguzmann89@gmail.com
Laura Becker /laura bɛkɐ/, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

languages in a way that the languages are as independent from each other as possible. An early example showing the need for bias control is found in Dryer (1989), who correctly identifies the difficulties which can arise from genealogical bias in a typological sample:

> What we would like to do is determine whether the difference between the frequencies of SOV and SVO order is statistically significant. But we cannot apply the relevant statistical tests, at least straight-forwardly, to Tomlin's data, because such tests require that the items in the sample be independent. But many of the languages in Tomlin's sample are not independent. As noted, for example, his sample contains 33 Bantu languages. In other words, although Tomlin's methodology allows one to obtain reliable estimates of the relative frequency of different language types among the languages of the world, it does not allow one to determine the extent to which those frequencies are due to linguistic factors, as opposed to nonlinguistic ones, and hence no way to determine whether there are statistically significant linguistic preferences for one language type over another. (Dryer 1989: 261)

This quote reflects a common worry in typology: oversampling. Put differently, we should not include too many languages from a specific family because doing so would lead to bias towards that particular family. Bentz et al. (2015) capture the idea of systematic sampling being the solution to reduce bias in the following way:

> Typologists know it is crucial to control for the non-independences in a dataset that stem from language areas and language families (e.g., Dryer, 1989, 1992). The best remedy for an areally and genealogically biased typological analysis is to balance the sample with respect to families and areas. (Bentz et al. 2015: 19)

However, the role of sampling in typology and its effectiveness at reducing bias has not stayed unchallenged and has been called into question by e.g. Himmelmann (2000) and Cysouw (2011). In this paper, we reassess both positions and argue that strict systematic sampling is not necessary to control for areal and genealogical biases. If the aim of a study is to examine crosslinguistic tendencies, we argue that some form of bias control is necessary. Using modern statistical techniques, we show that it is possible to control for genealogical and areal effects more effectively. Most importantly, though, statistical bias control does not require us to exclude languages from the sample that may be available otherwise. We also argue that the role of language contact in shaping linguistic patterns might be much stronger than generally assumed. To discuss the statistical methods of bias control, we focus on one specific question, namely the relation between verb-object order and affix position (prefixation vs. suffixation).

The paper is structured as follows. Section 2 presents the linguistic background on the relation between word order and affix position. Section 3 gives an overview of previous approaches to sampling. We then present our approach to

this issue in Section 4, proposing two statistical methods as alternatives to bias control in the sampling process. Section 5 discusses the results of our model. In order to relate our results to other common modeling choices in typology, we compare our model to two alternative model specifications and test for the robustness of our model to oversampling in Section 6. Section 7 discusses the findings of the previous sections and relates them to the theoretical discussions concerning sampling and bias control in language typology. Section 8 concludes the paper.

2 Word order and affix position

The observation that OV-languages strongly prefer suffixation while VO-languages show both prefixation and suffixation preferences has widely been discussed in the literature, e.g. in Bybee et al. (1990), Cutler et al. (1985), Dryer (1992), Hawkins and Gilligan (1988), and Siewierska and Bakker (1996) (see Song 2012: 54–65 for an overview). There is no final consensus about the reasons for the attested preferences, but we can distinguish two main approaches to account for this word order and affixation pattern: processing ease or efficiency and the diachronic processes leading to the distribution.

The processing explanations take the crosslinguistic distributions of patterns to reflect processing ease for the speaker: more frequent patterns are argued to be more easily or efficiently processed, and rarer patterns are regarded as more difficult to process for the speaker. In this vein, Cutler et al. (1985), Hawkins and Cutler (1988), and Hawkins and Gilligan (1988) apply the *Head Ordering Principle (HOP)* to affixation in that affixes are taken to be the heads of the lexical material in the same way that verbs are of e.g. objects. The *HOP* then states that languages prefer to have consistent head-dependent orders, i.e. that affixes and their lexical hosts occur in the same order as the verb and the object. The authors argue that this principle can functionally motivate the preference of OV-languages to occur with suffixes. However, according to the *HOP*, VO-languages should have a strong prefixation preference. While VO-languages certainly occur with prefixes more frequently than OV-languages, the distributions suggest that they equally take suffixes. To account for this additional compatibility with suffixation, the authors argue for a general suffixation preference based on processing considerations as well, which does not depend on the word order type. For VO-languages, both principles predict a preference for suffixes, as is attested in various samples. For OV-languages, on the other hand, the two principles are assumed to be in competition with each other, leading to more flexibility of OV-languages to occur with both suffixes and prefixes.

Other authors (Bybee et al. 1990; Dryer 1992; Siewierska and Bakker 1996) have argued that the preferences we can observe in crosslinguistic samples are rather the outcome of the processes leading to them and cannot be motivated on synchronic grounds.[2] Thus, the preference for e.g. tense suffixes over prefixes in OV-languages can be accounted for by the recurrent pattern that auxiliary verbs, a common source for tense affixes, tend to follow the main verb in OV-languages (Bybee et al. 1990: 9). The strong preference for suffixation over prefixation in OV-languages should then, at least partially, be due to the fact that the source elements often already follow the verb instead of preceding it, i.e. the correlation between verb-final order and affix position is not independent of the order of main and auxiliary verbs.

Bybee et al. (1990) show that affixes expressing verbal categories systematically occur in the position in which their source element appeared in, making a synchronic processing-based account unnecessary. The authors also point out that the observable distribution of the suffixation preference could be the result of two distinct constraints on diachronic processes. It could be the case that grammatical material generally tends to follow lexical material, resulting in the more frequent development of suffixes than prefixes. Or it could be that the order of lexical and grammatical material has no overall preference, but that preposed grammatical material develops into affixes less frequently than postposed grammatical material does.

Bybee et al. (1990) also show that certain semantic factors are relevant for whether or not a potential affix-candidate will fuse with the verb or with other available hosts. In verb-initial or verb-final constellations, affix-candidates often appear in utterance-initial or utterance-final positions, respectively, which means that in such contexts, the verb is often the only potential host. In verb-medial orders, on the other hand, affix-candidates do normally not occur in such edge positions and thus have an alternative adjacent potential host in addition to the verb. Bybee et al. (1990: 29–34) find that functions that are semantically more relevant to the verb itself, e.g. valency, which can form a more coherent conceptual unit with the verbal stem, are more frequently expressed as prefixes on the verb. On the other hand, markers of functions that are more relevant to the proposition itself such as mood/modality are less likely to fuse with the verb and become verbal prefixes. In the verb-medial languages of their sample, valency markers preceding

2 Himmelmann (2014) argues against such a synchronic, functional explanation of the overall suffixation preference as well. Focusing on affixation rates in preposed versus postposed grammatical material, he proposes a prosody-based account for the higher rate of affixation of postposed grammatical markers.

the verb were prefixes 73% of the time, while only 10% of preverbal mood/modality markers were prefixed (Bybee et al. 1990: 31).

A related issue was pointed out by Dryer (1992), i.e. that crosslinguistic tendencies of affix positions greatly depend on the specific types of affixes. In the sample used in Dryer (1992), tense or aspect markers on verbs show the "expected" behavior and tend to be suffixed in OV-languages and appear as both suffix or prefix in VO-languages. Nominal possessive markers, however, are slightly more likely to be prefixed in OV-languages but not in VO-languages (Dryer 1992: 127).

Siewierska and Bakker (1996) look at the association between word order and affix position, in this case focusing on subject and object agreement markers on verbs. Siewierska and Bakker (1996: 129–136) compare their findings in detail against previous typological studies concerned with the relation between word order type and the availability (and position) of affixal subject agreement markers (Foster and Hofling 1987; Hawkins and Gilligan 1988; Nichols 1992; Steele 1978), noting a number of divergences in the distributions across seven samples, some of which Siewierska and Bakker account for by different areal biases present to a varying extent in each of the samples.

For instance, Nichols's sample contains the highest proportion of subject agreement markers while Stassen's sample (cf. Appendix I in Hawkins and Gilligan 1988) shows the lowest of seven different samples that Siewierska and Bakker (1996) compare. At the same time, Nichols's sample contains the highest proportion of North American languages, while Stassen's sample has the lowest. The latter sample also only includes a small number of languages from Papunesia and Australia, which are areas that are identified as having high proportions of subject agreement markers across all word orders in the other samples (Siewierska and Bakker 1996: 130). Regarding such biases due to differences in sampling, Siewierska and Bakker (1996: 152) note: "[t]he questions concerning sampling methodology which this investigation has raised cannot be easily resolved. An awareness of the issue, however, is crucial for an appreciation of the nature of typological claims and the validity of the argumentation based on insights from typological research". This illustrates that even though typologists have generally been aware of the potential biases due to sampling methods and of how such biases can affect their findings, it still is an important issue deserving our attention.

Finally, work on verbal affixes by Enrique-Arias (2002) has suggested that the preference for suffixes over prefixes may be entirely due to other grammatical factors rather than basic word order. Enrique-Arias notes that subject and object agreement markers do not seem to follow the same patterns as other affixes, and he suggests that the suffixing preference is not directly related to OV or VO orders, but that it results instead from the combination of grammaticalization paths and

psychological factors, namely that speakers tend to resist the fusion of prefixing material but favor the fusion of suffixing material.

In summary, the question of the relation between verb-object order and affix position is still relatively open. We are not aware of a more recent study that has investigated this association with a special focus on the role of sampling; therefore, this question provides an optimal testing ground for the methods proposed in this study. As we will show in Section 5, our findings point towards little to no association between VO-orders and affix position, once genealogical and areal biases are controlled for.

3 Sampling methods in language typology

Creating representative language samples is an essential task in typology, in order to examine crosslinguistic distributions of patterns and to draw generalizations about language as such. Since sampling is of such importance, we find a substantial body of literature whose discussion would go beyond the purposes of this paper. In this section, we will only provide a brief overview of some sampling methods that are relevant to the understanding of the motivation of the present paper. More detailed overviews about sampling are provided in e.g. Bakker (2010), Miestamo et al. (2016) and Song (2018).

3.1 Probability and variety sampling

There are two main types of samples, each of which is designed for different purposes: probability sampling and variety sampling. Probability sampling is used to examine the crosslinguistic distribution of a certain feature, with the languages in the sample being as independent from each other as possible, so that inferences about crosslinguistic tendencies can be drawn. This type of sampling is often applied when the values of the feature at hand are already known, and when the research question focuses on the underlying crosslinguistic distribution of those values.

Variety sampling, on the other hand, serves the purpose of exploring the value space of a given linguistic feature, often at the initial stages of a research project. According to Miestamo et al. (2016: 234), the goal of variety sampling is to "display as much variety as possible in the linguistic realizations of the phenomena under investigation and to reveal even the rarest strategies or types of expression in the domain explored." A good case in point is mentioned by Song (2018: 84), who notes that "[i]t had been widely believed that there were no object-initial languages

in the world until the late 1970s when object-initial languages (spoken mainly in the Amazon) began to be brought to the attention of the wider linguistic community".

Often, however, samples are, at least to a certain extent, convenience samples restricted by the languages for which the relevant information is available to the linguist. Related to such practical issues of compiling a proper sample that fulfills the theoretical requirements, Himmelmann (2000) argues against a strict sampling method as a general requirement for all typological studies. Himmelmann (2000: 9–11) emphasizes the point that in order to find out about the types of patterns available, adding more languages to obtain a larger sample is more important than controlling for genealogical and areal independence in a restricting way. On the other hand, he also argues that at the initial stages of exploring possible structures, a smaller sample may be sufficient to put forth an initial hypothesis that can then be taken up and examined further by the linguistic community. Thus, Himmelmann emphasizes that controlled sampling can but does not have to be of high importance for "just any" study in language typology:

> More recently, however, sampling procedures have been elevated to the status of a general measure of the quality of typological work, regardless of whether or not it involves claims about frequency. This view, which is most explicitly expressed in recent reviews of typological books, seems to me to be ill-advised. Explicit sampling procedures are not relevant for just any kind of typological work. (Himmelmann 2000: 10)

Similarly, Cysouw (2011) calls into question the effectiveness of sampling, identifying the main issues with sampling as a method for controlling for autocorrelations:

> First, sampling reduces the already rather limited amount of data available about the world's languages, so any generalisation has to be made on the basis of less than possible data […] Second, there might be unrecognized genealogical or areal groupings, not acknowledged in the sampling, which leads to inflation of the frequency of a type, notwithstanding the sample. Even more problematic is the possibility that the actual world's languages are not representative of the possible human languages. (Cysouw 2011: 416)

3.2 Types of biases

Since probability samples are used to obtain a representative selection of the languages of the world, they are built to overcome potential biases as much as possible. This is crucial because in order to generalize from a language sample to universal tendencies, each language, being a data point, needs to be as independent as possible from other languages (or, as we will see, the bias needs to be controlled for in the modeling). Bakker (2010) and Song (2018) distinguish the

following types of bias that can distort a language sample: bibliographical, genealogical, areal/contact, typological, and cultural.

While the bibliographical bias is a practical issue, it is an important one. Bakker (2010: 91) notes that "over two thirds of the known languages have not been described at any level of linguistic sophistication"; Song (2018: 78) gives the proportion of "less than 10%"; and Hammarström et al. (2020) indicate that long grammars (>300 pages) are only available for about 23% of the world's languages.

Most theoretical discussions of bias in language samples focus on genealogical and areal or contact biases. If two languages have a common ancestor, or have been in contact with each other, finding that they share their value for a linguistic feature cannot be interpreted as two independent data points sharing that feature. Thus, if genealogical or areal biases are not taken into account, crosslinguistic generalizations run the risk of being biased by a linguistic area or a family. Dryer (1989: 259–261) makes this point for genealogical bias in word order preferences:

> Secondly, […] about 40% of the SVO languages in the world are Niger-Congo languages. If it were not for whatever historical factors led to the large size of this family, particularly those leading to the relatively recent expansion of speakers of Bantu languages, the number of SVO languages in the world would have been considerably lower, in fact not much more than half the number of SOV languages. (Dryer 1989: 260)

Similarly to genealogical bias, areal bias can occur in a sample with various languages that are or have been in contact or that are part of a larger area in which the same value of a linguistic feature has spread by language contact and diffusion over time. Such linguistically relevant areas can be of various sizes; areas known to include languages that are not related but still share certain properties due to contact are e.g. the Balkans, the Baltic, Ethiopian highlands, South Asia, the Sepik River basin (New Guinea) and the Pacific Northwest of North America (Thomason 2001: Chapter 5).

A so-called typological bias arises when the association between two features is examined with a biased distribution of one of the values. To illustrate this issue, Dryer (1989) refers to the results from Nichols (1986), which show a stronger preference for verb-initial languages to have head marking than for other word order types. Dryer (1989: 264–265) observes that out of the 13 languages in her sample classified as head marking, 10 languages are spoken in North America (including Central America). Also, the three languages that are verb-initial but not head marking are all spoken outside of North America. Head marking being much more frequent in North America than in the other areas of the world, "it would appear that the supposed association between head-marking type and verb-initial order is simply an artifact of the fact that most of the verb-initial languages in Nichols' sample are from North America and the fact that the head-marking type is

considerably more common in North America than it is elsewhere in the world" (Dryer 1989: 265). We will show in Section 5 that the situation of verb-object order and affixation is similar. Once genealogical and areal bias is controlled for, we do not find any clear effect of verb-object order on affixation preference.

Cultural bias is another type of bias that is strongly intertwined with genealogical and areal biases. In a stricter sense, a cultural bias can occur when languages with a shared or similar cultural background may be biased for or against certain linguistic structures in the same way. In a broader sense, cultural bias also includes the socio-linguistic properties of the speaker groups and their impact on grammar. It has been shown that crosslinguistically rare patterns tend to occur in languages with smaller speaker communities. For instance, Nettle (1999: 133–134) shows that the median speaker community size of rare OVS and OSV word order is 750, whereas the median community size of all human languages is approximately 5,000.

3.3 Previous sampling methods

There have been numerous proposals for building language samples (Bell 1978; Bickel 2008; Dahl 2008; Dryer 1989, 1991, 2018; Miestamo et al. 2016; Nichols 1992; Perkins 1980, 1989; Rijkhoff and Bakker 1998; Stassen 1985). We will focus on a number of important methods here (for more details on sampling methods, see e.g. Bakker 2010; Miestamo et al. 2016; Song 2018).

The earliest systematic probability sample is proposed by Bell (1978), who tries to build a language sample while avoiding genealogical bias by over-representing certain families. The procedure starts out with a classification into 478 groups of families. He takes the number of sub-groups as an approximation of linguistic diversity within the group, and includes one language per sub-group. The final number of languages from that group is then proportionally adjusted from 478 groups to the number of languages the final sample should contain.

Perkins (1980) elaborates on the method proposed in Bell (1978), adding emphasis on the cultural independence of the languages in the sample, combining a genealogical classification of languages (Voegelin and Voegelin 1977) with a cultural classification (Murdock 1967).[3] This results in a sample of 50 languages that are assumed to be genealogically and culturally as independent from each other as possible.

3 A refined description of his method is presented in Perkins (1989). Bybee et al. (1994) is another, further developed, application of this sampling method, resulting in a stratified probability sample.

Dryer (1989) also offers a refined sampling procedure of Bell (1978), using data on word order preferences from 542 languages. The first important innovation is that he uses genera as the main genealogical grouping, the idea being that genera are comparable in terms of time-depth (3,500–4,000 years). Dryer also samples genera instead of single languages. The second important innovation concerns the distribution of genera into 5 continent-sized macroareas.[4] The division into such large areas is not meant to prove the existence of linguistically relevant areas in the size of continents (Dryer 1989: 267); rather, the idea behind this is to divide the globe into areas which are arguably independent enough from each other due to physical boundaries preventing the expansion of languages, language contact, and diffusion to a reasonable extent. Dryer (1989) argues that only if a certain pattern or association can be shown to exist in all (independent) macroareas, can we safely exclude genealogical or areal bias and draw conclusions about general crosslinguistic tendencies. To this end, he proposes to count the genera in each macroarea that exhibit the relevant linguistic pattern. In case a genus contains languages that exhibit more than one pattern (e.g. OV as well as VO-orders), this genus is counted twice, i.e. into both patterns (Dryer 1989: 289). After establishing the proportional distributions of all relevant patterns in each macroarea, the latter are tested separately: "The final step is to determine how many of the five areas conform to the hypothesis being tested. If all five conform, then the hypothesis is considered to be confirmed" (Dryer 1989: 269).

Nichols (1992) builds her sample in a similar way to Dryer (1989), distinguishing 10 "sample areas" that are taken to be geographical and cultural units.[5] She further distinguishes "stocks" with a time-depth of 5,000–8,000 years and families with a time-depth of 2,500–4,000 years (Nichols 1992: 24–25). All stocks and families are represented in the sample by a single language (as far as possible), with large and diverse families such as Indo-European being restricted to include six languages. In contrast to Dryer (1989), Nichols samples actual languages instead of genera.

---

4 Dryer (1989) originally distinguished Africa, Eurasia, Australia-New Guinea, North America, and South America. In Dryer (1992), he proposed a 6 macroarea distinction, which was later refined in Hammarström and Donohue (2014) and Dryer and Haspelmath (2013), leading to the by now commonly used 6 macroareas of Africa, Eurasia, North America, South America, Papunesia, and Australia.

5 The 10 areas are Africa, Ancient Near East, Northern Eurasia, South and Southeast Asia, New Guinea, Australia, Oceania, North America, Mesoamerica, and South America. This classification was later refined in the AUTOTYP project (Bickel and Nichols 2013); it contains two areal configurations, one with the continent-size areas suggested by (Nichols 1992), and one with these areas further subdivided into 24 smaller areas.

The sampling methods described so far all build genealogically as well as areally and culturally stratified probability samples, but they do not consider the linguistic feature(s) in question for the sampling process itself. There is an alternative way of arriving at probability samples, namely a posthoc sample that departs from a larger sample of languages and takes into account the distribution of linguistic feature(s) in order to select languages. Dahl (2001, 2008) and Bickel (2008) are just such approaches.

Dahl (2008) proposes a method for creating samples of languages based on how typologically similar they are according to WALS (Dryer and Haspelmath 2013), i.e. the sample is linguistically informed and takes into account how similar languages are across a number of linguistic features. The typological distance (or similarity) between two languages is defined as the proportion of features with different values in those languages (Dahl 2008: 211).

Another posthoc method of sampling is provided by Bickel (2008). The main idea of controlling for genealogical bias is similar to the one in Dryer (1989) in that it also builds a sample of genera (or any other level of genealogical group) instead of using single languages. However, Bickel (2008: 223) notes that the assumption in Dryer (1989) that languages from the same genus will mostly display homogeneous properties is problematic, since this is not necessarily the case. Moreover, whether or not a linguistic feature is stable in a genus across time also provides useful information and should be taken into account when examining crosslinguistic trends. Thus, Bickel (2008) exhibits two important innovations. First, he uses the entire phylogenetic information available, including more than one genealogical level. Second, he takes into account the value distributions of the feature in question to determine how many related languages (on any level) can be included in the sample.

The algorithm proposed in Bickel (2008, 2011) starts out from the highest grouping of languages and determines whether or not the expression of the feature in question is skewed towards a certain value in a group of related languages or not. If the values (e.g. VO vs. OV-orders) are approximately equally represented in a family, Bickel (2008: 224) argues that "it is likely that genealogical membership is irrelevant for the distribution of values in the unit, and there is no reason to include only one datapoint per distinct value in the sample". If the distribution of values is skewed, e.g. 90% versus 10%, "it is likely (though by no means necessary!) that the distribution is induced by shared retention, innovation or family-bound drift – i.e. it is a skewing that we want to control for" (Bickel 2008: 224). Then, this genealogical group will contribute only one datapoint with the majority value to the sample in order to prevent a bias towards that value. The minority value, however, if present in more than one language, is counted in with all datapoints. Bickel (2008: 224) motivates this in the following way: "the presence of the minority value

must be due to some non-genealogical factor, i.e. perhaps it was precisely one of the areal or structural factor under investigation that triggered the deviation." This process is repeated for each level of genealogical groups, from the highest to the lowest one.

All methods presented so far are probability sampling methods. We now turn to the two most important variety sampling methods (i.e. sampling with the aim of capturing the linguistic diversity as much as possible) are those by Rijkhoff and Bakker (1998), Rijkhoff et al. (1993), and by Miestamo (2005) and Miestamo et al. (2016).

The method developed by Rijkhoff and Bakker (1998) and Rijkhoff et al. (1993), called "Diversity Value" (DV) is a fairly involved but formalized process to select an optimal number of languages on the basis of any given hierarchical structure of language. Thus, for e.g. phylogenetic stratification, it requires a full tree-structure considering all genealogical groups and the branching structure within groups. In DV sampling, the number of languages from each genealogical group should be proportional to the amount of variation (i.e. the "Diversity Value") within that group so that the degree of linguistic variation is represented proportionally. The amount of variation is approximated by the number of all non-terminal nodes in a given genealogical group. Therefore, it is the structure of the genealogical tree (or any other kind of taxonomic tree), rather than the number of languages per group, which determines the number of languages included per group. In addition, all major groups are required to be represented by at least one language. What sets this method apart from the other ones described so far is that it does not include any direct areal stratification mechanism.

Another variety sampling method, called "Genus-Macroarea method" (GM) has been developed and described in Miestamo (2003, 2005) and Miestamo et al. (2016). It uses the classification into macroareas and genera according to the WALS (Dryer and Haspelmath 2013). The aim is to include languages from as many genera (Dryer and Haspelmath 2013) as possible in order to capture crosslinguistic variation in an adequate way. The first step leads to the genus sample of 523 languages, one from each 521 genera and one pidgin or creole and one sign language. Since it may practically be impossible to find a suitable language for each genus, the authors define the core sample to contain the maximal number of languages available for a given research question. Such a core sample will most likely contain bibliographical and thus areal biases; therefore, the authors suggest an areal stratification process to select a restricted sample from the core sample in a way so that all areas are proportionally represented by a number of languages according to the area that is most under-represented.

All these sampling methods share the general intuition that universal preferences or patterns can be captured by examining the synchronic crosslinguistic

distribution of a given feature (combination) in a sample, representing the languages of the world as adequately as possible. An entirely different approach to assessing universal preferences is to examine the transition probabilities across different values of a given feature over time. There are several proposals for how transition probabilities should be estimated (Dediu 2011; Dediu and Cysouw 2013; Dediu and Levinson 2012; Dunn et al. 2011; Levinson et al. 2011; Maslova 2000; Maslova and Nikitina 2007; Pagel 1994) and interpreted (Cysouw 2011). We will not discuss such "dynamic" approaches in detail in this paper, since they re-frame the question of how universals should be studied and formulated from a synchronic distribution to the probabilities of diachronic changes (but see Section 7.2 for a comparison of "static" and "dynamic" approaches).

However, there is a crucial conceptual point in using transition probabilities that is also inherently part of the approach to bias control that we propose in this paper. Any method of estimating transition probabilities requires a sample of related languages for the estimation of how likely a feature is to change its value over the course of time. Thus, what would be treated as a potentially problematic bias for the other sampling methods mentioned above is an integral part in estimating transition probabilities. In other words, they require the information contained in the genealogical relation between languages in the sample. We show that this idea of using the information about genealogical (and also areal) relations between languages in a sample can also be applied to "static" sampling approaches and need not lead to the exclusion of languages from a sample.[6]

3.4 Towards more inclusive sampling

All sampling methods mentioned in the previous section include some mechanism of stratification to circumvent areal, genetic, or cultural bias. In practice, this always means discarding languages for which sources would be available. In the case of the two posthoc methods proposed in Dahl (2008) and Bickel (2008), the exclusion of data is even more rampant, since a potentially large portion of the data that had already been gathered will be excluded at a later step of the sampling process.

6 The g-sampling approach in Bickel (2008) is somewhat similar in that respect, as it also establishes criteria according to which more than one language from a genealogical unit can be included in the sample. However, if there are sufficient grounds to assume a strong genealogical bias for a given value of the feature at hand, the approach still falls back on including only a single language (or g-unit) from that group.

Strict genealogical stratification methods are all based on the default assumption that closely related languages must be similar because of the common source of the value for the relevant feature, regardless of which area of grammar that feature belongs to. Cysouw (2005: 556–557) raises the important theoretical point that this may not even be the case, mentioning the high degree of variation for indefinite pronouns in Romance and Germanic. He also notes that "[i]nstead of sampling one language per genealogical unit, it is actually much more informative to sample various languages from the same unit" (Cysouw 2011: 421). By now, we have the tools to build in the degree of dependency between languages into a statistical model, so that we no longer need to exclude available datapoints (see Section 4 for our proposal). In addition, including languages that are closely related, or in close contact with each other, can also help to test whether or not a given feature is realized similarly. If many languages from a single genealogical group express a feature by the same inherited value, this is also meaningful information that tells us something about the prevalence of that feature value. Moreover, in some cases where diachronic data may not be available, we may not even know for certain that a feature shared between related languages spoken in close areas necessarily goes back to the common ancestor language; it may have spread by contact or areal diffusion as well. Including closely related languages also plays a very important role for examining an implicational universal (if a language has X, then it also has Y): high uniformity within a family may be indicative of a very stable association.

4 Our approach

4.1 Dataset

The dataset used in the present study is taken from WALS chapters 26 and 83 on "Prefixing versus suffixing in inflectional morphology" and "Order of object and verb", respectively (Dryer 2013a, 2013b). It includes the intersection of languages for which data from both chapters is available. This amounts to a total of 780 languages across 158 macrofamilies.[7] Figure 1 shows the geographic distribution of the languages in the sample together with their affix position and verb-object

7 We compiled the data on December 12 2019, and did some minor manual clean ups. The dataset as well as the code are provided in the Supplementary Materials (https://doi.org/10.5281/zenodo.5576242).

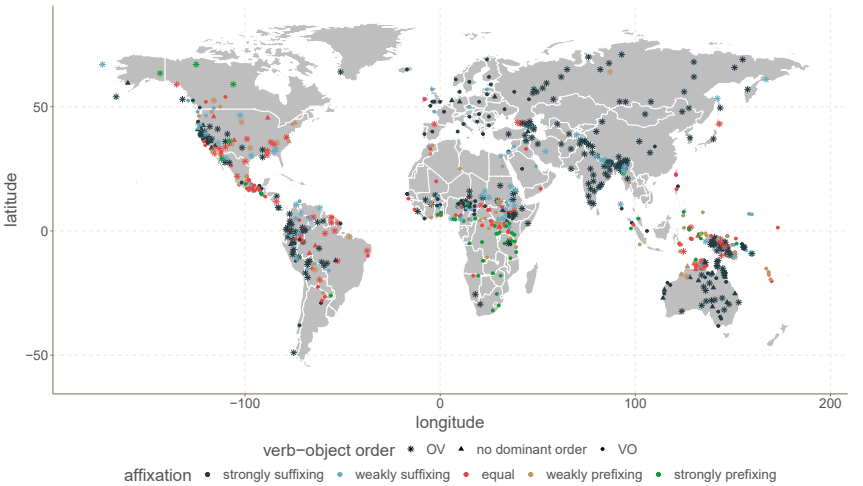

Figure 1: Distribution of languages in the sample.

order. Impressionistically, we find some regions which seem to prefer prefixation over suffixation, e.g. Oceania, North America, and the Southern parts of Africa. The same holds for verb-object orders; Western Europe and Central Africa tend to have VO-orders, while languages in Asia are predominantly OV.

Concerning affix position, Dryer (2013b) distinguishes six different values according to which each language is classified, out of which we include the five types that show inflectional morphology: predominantly suffixing (strongly suffixing), moderate preference for suffixing (weakly suffixing), approximately equal amounts of suffixing and prefixing (equal), moderate preference for prefixing (weakly prefixing), predominantly prefixing (strongly prefixing).[8] In his classification, Dryer (2013b) distinguishes 10 types of affixes: case affixes on nouns, pronominal subject affixes on verbs, tense-aspect affixes on verbs, plural affixes on nouns, pronominal possessive affixes on nouns, definite or indefinite affixes on nouns, pronominal object affixes on verbs, negative affixes on verbs, interrogative affixes on verbs, and adverbial subordinator affixes on verbs. He then assigns each language a suffixing and prefixing index.[9]

8 We excluded the 140 languages of the type "little or no inflectional morphology". Since we use the family tree information in Glottolog (Hammarström et al. 2020), we also excluded languages with no information in Glottolog (5 in total).

9 The exact manner in which Dryer calculates the prefixing and suffixing indices is rather complex. For more details, see Chapter 26 in WALS (Dryer 2013b).

Languages with a suffixing index over 80% are classified as predominantly suffixing, indexes between 60% and 80% are treated as a moderate preference for suffixing, indexes between 40% and 60% as approximately equal amounts of suffixing and prefixing. Finally, languages with a prefixing index between 60% and 80% are classified as having a moderate preference for prefixing, and prefixing indexes over 80% are treated as predominantly prefixing.

Figure 2 shows the distribution of those five affixation types in the dataset, and we can see their distribution across macroareas in Figure 3. Overall, most languages are strongly suffixing and the strongly prefixing type is very rare. However, this trend is most pronounced in Eurasia and Australia, while the other macroareas show a much weaker preference for suffixation. Africa, Papunesia, and North America are the macroareas with the highest proportion of languages with a strong or weak prefixation preference.

The word order types distinguished in Dryer (2013a) are OV, VO, and no dominant word order. Their distribution in the dataset across macroareas is shown in Figure 4. We can observe two main macroareal trends: all macroareas save for Africa show a preference for OV-order with Eurasia having the strongest trend, and having no dominant word order is comparatively frequent in Australia and North America.

Merging the datasets of Dryer (2013a, 2013b) allows us to examine the association between word order type and affix preference. Figure 5 shows the

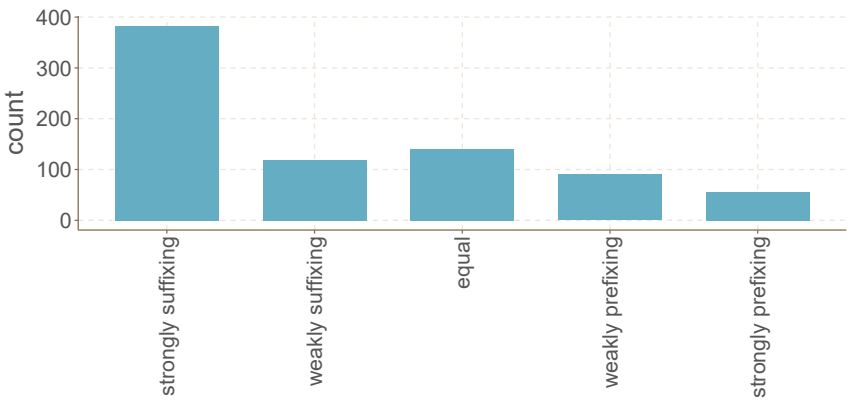

Figure 2: Overall distribution of affix position.

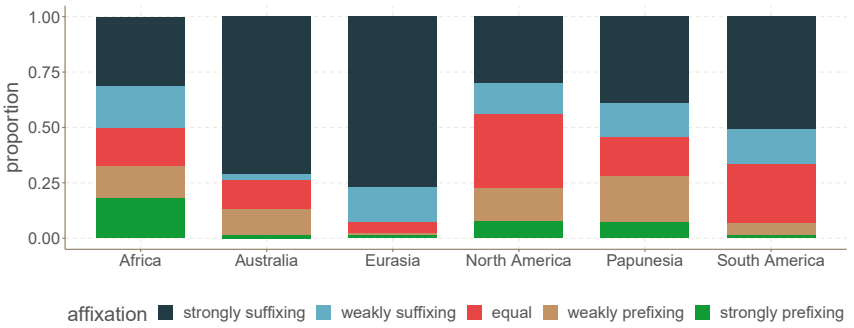

Figure 3: Affix position by macroarea.

distribution of verb-object orders across affix positions based on the dataset provided in Dryer (2013a, 2013b). We see that in raw numbers, OV-languages have a clear preference for being strongly suffixing, while VO-languages do not appear to be associated with a single value of affix position. VO-languages do however have a higher number of both weak and strong prefixation than OV-languages or languages without dominant word order. This distribution is in agreement with previous findings based on different samples (cf. Section 2). As we will see in Section 5, this apparent correlation is likely due to genealogical and areal bias in the sample.

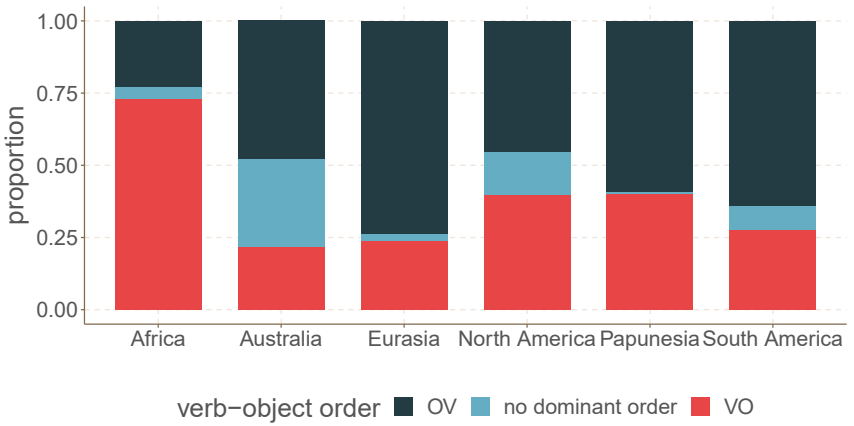

Figure 4: Distribution of word order by macroarea.

4.2 Bias control for genealogical effects

4.2.1 Previous approaches

As was pointed out in Section 3.3, all sampling methods are designed to control for genealogical bias in one way or another, and we can distinguish three types of approaches. The first, and probably most common one, is to control for it during the sampling process itself. The second one, developed by Dryer (1989) and adapted by Bickel (2008), is to sample genera instead of families, taking into account the variation within genera, or other levels of genealogical groupings. Finally, a more recent technique is building a statistical model and including the genus (or any other level of genealogical grouping) as a group-level effect in the model (Bentz and Winter 2013; Blasi et al. 2019; Jaeger et al. 2011; Levshina 2019).[10]

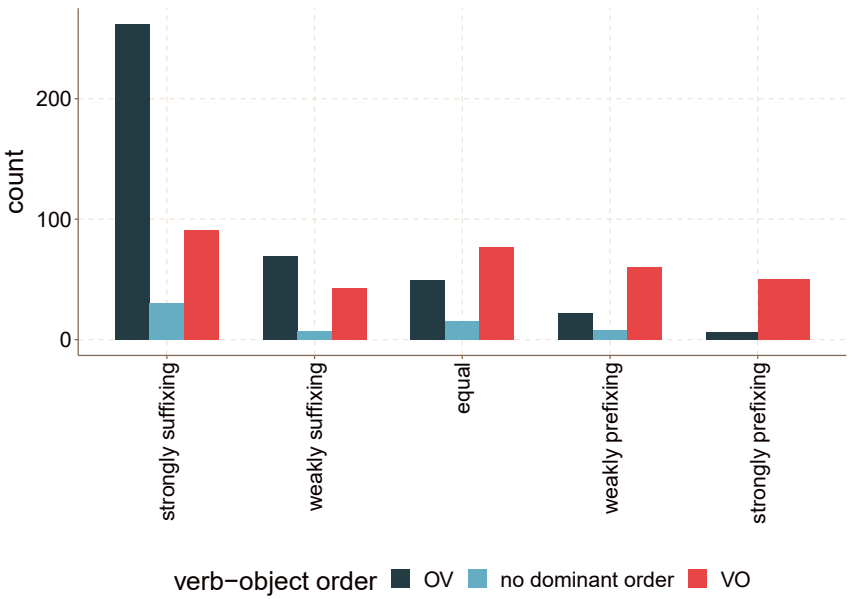

Figure 5: Verb-object order by affixation preference.

10 Group-level effects are also known as random effects in frequentist statistics.

However, including genera or families as a predictor in a model comes with similar potential issues as *a priori* restricting the number of languages from a genus in the sample. It is not necessarily known beforehand which level of genealogical grouping is most suited to examine a certain linguistic feature. There may be structure above and below it that is ignored, whichever grouping level is chosen. For instance, if one uses only language families, i.e. established groups of the highest level (e.g. Indo-European or Austronesian), lower-level groupings within families will be ignored by the model and it will not be able to capture that e.g. Germanic languages are expected to be more similar to each other than to Indo-Aryan languages. Moreover, it is not clear that all family or high-level groupings are of comparable generality and time depth. As mentioned in Section 3.3, Dryer (1989) established genera as a genealogical unit with a time depth of 3,500–4,000 years to account for this issue, and with its use in the WALS database, this level of grouping has become a common group-level effect in statistical models of typological studies. But even using comparable units such as genera as a predictor in a model, one may still miss important lower and higher-level relations.

This issue was already identified in Rijkhoff et al. (1993) and Rijkhoff and Bakker (1998), who pointed out that it is not only important to consider families but to include information about the entire phylogenetic tree, which the authors implemented in their method of Diversity Value sampling.

Other approaches that are designed to include several levels of genealogical groupings are presented in Bickel (2008) (g-sampling) and Bickel (2015) (family bias method). As explained in Section 3.3, g-sampling involves recursively checking for bias in genealogical groupings from the highest to the lowest defined level, building a sample of those genealogical units which are assumed to be genetically unbiased. This method, however, does not allow construction of a sample of individual languages that could be used to address several typological questions but only provides a sample of genealogical units that are useful for a specific research question.[11]

Bickel's family bias method tries to estimate whether there is a preference for a particular feature value within a language family. Note that in contrast to g-sampling, this method does not aim at building a balanced sample but at, very indirectly, estimating the transition probabilities between different values of linguistic features.[12] This is independent of whether the preferred value was already present in the proto-language or not, and whether the synchronic distribution

11 See Bakker (2010) for a similar comment about the method of Dryer (1989).

12 A more direct approach to estimating transitional probabilities is presented by Maslova (2000), Maslova and Nikitina (2007), and Jäger and Wahle (Forthcoming).

is the result of innovation or preservation. The method can easily be applied to groupings with a sufficient number of languages (Bickel suggests at least six, but this depends on the desired degree of certainty). For groupings with fewer or only one member, Bickel (2015) suggests extrapolation from families with more members.

4.2.2 Our approach: phylogenetic regression

In this study, we propose to use hierarchical phylogenetic regression (de Villemereuil and Nakagawa 2014: ch. 11; Garland and Ives 2000; Housworth et al. 2004) as a way of controlling for family bias.[13] Including a phylogenetic term in a regression model is conceptually very similar to including a categorical group-level effect such as genus. When including genus as a group-level effect, we assume that languages from the same genus are very likely to share a given linguistic feature. The phylogenetic term also controls for such genealogical relations, but instead of "bagging" languages into genealogical groups to which a language either does or does not belong, it takes into account the entire phylogenetic tree. Doing so allows us to represent the assumption of similarity between languages due to their genealogical relatedness in a gradient way.

To build the phylogenetic term, we used the information on family trees from Glottolog 4.3 (Hammarström et al. 2020). We chose Glottolog because, as far as we are aware, it is the most complete and exhaustive genealogy available at this moment, subjected to peer-review and updated regularly. Of course, it is possible to use any other language tree. For practical reasons, we use "micro-families" as the smallest unit. We define micro-families as the smallest genealogical groupings above the leaf nodes in the trees provided by Glottolog.[14] For the most part, micro-families include only a single language, but there are some which can include two or three very closely related varieties of languages. For instance, the micro-family of Spanish is Castilic and the micro-family of German

13 Hierarchical regression is also known as mixed-model in the frequentist literature. Phylogenetic regression should not be confused with phylogenetic models used for building phylogenetic trees (or networks) to establish or examine the relations between languages (e.g. Bouckaert et al. 2018; Bowern and Atkinson 2012; Gray et al. 2009; Jäger 2013, 2018, 2019; List et al. 2014; Murawaki 2015, 2018; Verkerk 2019), or for assessing the stability of grammatical features over time (Dediu 2011; Dediu and Levinson 2012; Dunn et al. 2011; Maurits and Griffiths 2014; Murawaki and Yamauchi 2018).

14 Leaf nodes in the Glottolog trees can but do not have to correspond to the language level in the WALS dataset.

is Global German. Using micro-families instead of single languages was a practical choice; building a phylogenetic tree with single languages did not improve the models but only made them slower, harder to fit, and led to overfitting.

To build the phylogenetic tree, we first determined the distance, i.e. similarity, between all pairs of micro-families. The similarity between two micro-families was calculated as the total number of shared nodes in the Glottolog tree between them. The resulting similarity matrix was then used to build a phylogenetic tree for the languages in the sample. From this tree, we calculated a phylogenetic covariance matrix for all micro-families, which reflects the hierarchical structure between all micro-families.[15]

Table 1 shows the covariance matrix for a subset of micro-families. With values ranging from 0 to 1, values closer to 1 mean that the languages are very similar, while very distantly related or unrelated languages have values closer or equal to 0. For instance, the covariance matrix captures the fact that Italian, French, and Spanish (Castilic) are much more similar to each other, having covariance values of 0.9 and 0.91, than they are to Germanic (German and Dutch) or Indo-Iranian (Hindustani), with a covariance of 0.67. Two languages which are entirely unrelated to each other have a covariance value of 0 in the matrix. Table 1 shows this with Warrwa (Nyulnyulan, Australia) and Yatê (Fulniô, South America), neither of which belong to the Indo-European family. In this case, we do not distinguish between isolates (e.g. Yatê), which do not have other known related languages, and languages that are the only representative of a given family in our dataset (e.g. Warrwa).

This covariance matrix is then included in the model in the following way: we add a group-level effect to the model for each micro-family, enforcing the correlation between the intercepts for micro-families to follow the covariance matrix.[16,17] Values closer to 1 mean that the intercepts of the micro-families will be

15 See the Supplementary Materials for the code for building the phylogenetic tree.

16 More precisely, a hierarchical model including a phylogenetic term has the form $y = \mu + \beta x + a + \epsilon$, where $\mu = \beta x$ are the intercept and coefficients for the covariates, and a is the term for the phylogenetic effects with $a \sim N(0, \sigma_P^2\,\Sigma)$. Σ is the phylogenetic covariance matrix, and ϵ is the residual error: $\epsilon \sim N(0, \sigma_R^2\,\mathrm{I})$. This mean that σ_P^2 is the variance of the phylogenetic effect and σ_R^2 is the variance of the residual error.

17 In principle, one can also add varying slopes to the model. However, this makes the model computationally very challenging. With our dataset and the hardware that we have access to (a High Performance Computing server with 4 Intel Xeon Gold 6140 processors, 144 cores, and 754 GB RAM), we were unable to fit such a model within a reasonable period of time because each model would take upwards of two months to fit (we terminated the process after a month and the sampling was at 30%).

Table 1: Example covariance matrix for a subset of micro-families in the sample.

	Hindustani	Global German	Global Dutch	Castilic Spanish	Global French	Italian Romance	Fulniô	Nyulnyulan
Hindustani	1.00	0.67	0.67	0.67	0.67	0.67	0	0
Global German	0.67	1.00	0.83	0.67	0.67	0.67	0	0
Global Dutch	0.67	0.83	1.00	0.67	0.67	0.67	0	0
Castilic	0.67	0.67	0.67	1.00	0.91	0.90	0	0
Global French	0.67	0.67	0.67	0.91	1.00	0.90	0	0
Italian Romance	0.67	0.67	0.67	0.90	0.90	1.00	0	0
Fulniô	0	0	0	0	0	0	1.00	0
Nyulnyulan	0	0	0	0	0	0	0	1.00

estimated closer together in the model, and zero means that the intercepts can vary freely from each other. In other words, the model assumes the estimated effects to be more similar for micro-families which are closer together in the family tree, but less so for micro-families which are less close to each other in the tree. For the results of the phylogenetic effects in our model, see Section 5.3.

4.3 Bias control for areal and contact effects

4.3.1 Previous approaches

Areal and contact effects, besides being studied in related areas of linguistics, have always played an important role in linguistic typology. Such effects are relevant to broader, systematic crosslinguistic studies for a simple reason: languages that are spoken in (close) proximity to each other, and potentially by the same speaker community in a multilingual setting, influence each other as speakers borrow words, constructions, etc. from one language into another. A large body of work on language contact and borrowing over the last decades has shown that borrowing is not restricted to particular linguistic domains but can affect any part of grammar under certain circumstances (cf. Aikhenvald and Dixon 2006a, 2006b; Hickey 2010; Matras and Sakel 2008; Siemund and Kintana 2008; Thomason 2001 for overviews on language contact and borrowing). Language contact does not only affect neighboring languages, but it can lead to the diffusion of linguistic patterns across larger geographical areas over the course of time. While the genealogical

inheritance of a linguistic feature is often referred to as vertical transmission, contact and diffusion can be understood as horizontal transmission.[18] Since diachronic linguistic or socio-cultural data is often not available, both types of transmission are strongly intertwined and cannot always be distinguished from each other. Like genealogical relations between languages, contact and areal effects are important for systematic crosslinguistic studies because they represent a dependency between languages, and thus datapoints, in the sample. Particularly, Bickel (2017) emphasizes that patterns which may appear to have universal validity may mostly be due to areal effects:

> When we say that a structure (say, verb-final order) has spread in an area, what is meant is that the languages in this area changed their structure so as to mirror the structure of their neighbors, or that they selectively kept structures that mirror those of their neighbors. (Bickel 2017: 42)

Bickel (2017) argues for a distinction between functional triggers (e.g. cognitive preferences and constraints) and event-based triggers (effects of specific historical contingencies) of language change. Importantly, the former can be expected to have a similar effect across areas, while the latter should lead to linguistic features that are clustered in certain areas. According to Bickel (2013, 2015, 2017), typologists should aim to account for both types of triggers and to distinguish them from each other. At the same time, he notes that "we clearly need more 'meta-typological' research" to establish what the best methods are for modeling (different types of) areal effects (Bickel 2017: 45).

One of the most important proposals to systematically control for areal biases is made in Dryer (1989). As mentioned in Section 3.3, he divides the world into geographically independent macroareas and tests for tendencies within areas. Only if a tendency can be established with sufficient certainty in all macroareas, Dryer suggests, can a trend be assumed to be universal. Other important geographical stratification proposals are formulated by Nichols (1992), Bickel (2013), and Hammarström and Donohue (2014). The latter, building on Dryer (1989), propose a principled way of determining six macroareas that are as geographically independent from each other as possible and that are as comparable as possible in terms of their genealogical diversity.[19] Even though areal and

18 Note that representing contact-induced changes as horizontal transmission in opposition to genealogical inheritance or internal language change as vertical transmission is a strong simplification; language change always happens over the course of time and is therefore always "vertical" (cf. Croft et al. 2011: Section 4.4).

19 Other quantitative approaches to establish linguistic areas from the perspective of language typology are Bickel and Nichols (2006), Donohue and Whiting (2011), and Hammarström and Güldemann (2014).

contact effects have received substantial attention from the typological community as well (e.g. Aikhenvald and Dixon 2006a, 2006b; Bickel 2017; Enfield 2005; Hickey 2017; Holman et al. 2007; Matras and Sakel 2008; Nikolaev and Grossman 2018; Urban et al. 2019; and references therein), the focus of bias control in sampling still seems to lie more on genealogical than on areal stratification.

That areal effects have not necessarily been controlled for in typological work may partly be due to the fact that much of grammatical structure was long thought to be only inherited and not subject to borrowing or diffusion (cf. Enfield 2005), but it is certainly also due to the fact that language contact is still very difficult to quantify or operationalize (cf. List 2019).[20] We know that (structural) borrowing depends on many linguistic and extra-linguistic factors determining the contact situation.[21] Examples of important factors are the structural similarity of the languages involved, their political status, the level of multilingualism of the speakers, the intensity and duration of the contact, social structures (tightly-knit vs. more open), and language attitudes of the speaker communities.[22]

Even though it seems very difficult to generalize about language contact across languages and socio-linguistic settings, results of previous studies point to the non-negligible effect of contact. For instance, Holman et al. (2007) showed that for both related and unrelated languages, geographical distance is negatively correlated to structural similarity. In other words, there is spatial autocorrelation between languages. In a large dataset, languages that are geographically closer to each other generally also share more structural properties than with languages that are spoken at a larger geographical distance. Another finding that underlines the importance of areal controls in typological samples comes from Bickel and Nichols (2006). The study tests the hypothesis of the Pacific Rim as a large linguistic area and finds that, indeed, there is a weak signal for linguistic areality. The authors note:

> A troubling historical question: How could PR [Pacific Rim] variables persist so long in an area when there are many cases of their loss within historically reconstructed language families that are younger than the PR? Rather than a shortcoming we see this as a defining property of diagnostic areal features: they are more persistent in areas than in families. This must be because their retention can be favored by areal pressure, and because in linguistic areas they are prone to be transmitted not only by inheritance but also by substratal retention and diffusion. (Bickel and Nichols 2006: 7–8)

20 Regarding the test case of the present study, it is important to mention that "[c]lausal constituent order is highly susceptible to diffusion" (Aikhenvald 2006: 16).

21 See Aikhenvald (2006: 15–47) for an overview of the linguistic and extra-linguistic factors that influence borrowing and diffusion.

22 While structure borrowing has been shown to depend on the specific properties of a given contact situation, Seifart (2015) showed for affix borrowing that there is no evidence for it to be dependent on the structural similarity of the languages involved.

This suggests that areal effects may be expected to be as important as genealogical effects, at the very least. Thus, we have to assume that both affix position and word order may be subject to non-negligible areal and contact effects.

There are a number of ways in which quantitative typological studies have controlled for areal and contact bias. Many control for areal bias in the sampling process itself instead of in the statistical testing of associations or trends (e.g. Becker 2021; Hetterle 2015; van Lier 2016; Louagie and Verstraete 2016; Martowicz 2011; Miestamo 2005; Schmidtke-Bode 2009; Ye 2020).

Others, for instance, Donohue and Nichols (2011), Dryer (2011), and Sinnemäki (2014), test the strength of the association of linguistic variables in each defined area separately. Sinnemäki (2014) compares his results across macro areas and tests for significant differences in a second step. Other studies, e.g. Sinnemäki (2010) and Bickel (2011), include area as a predictor variable in a regression model together with the linguistic predictor(s) to model the outcome variable.

The important methodological papers of Cysouw (2010) and Jaeger et al. (2011) established the common practice of using hierarchical regression models in language typology with macroarea (and also family or genus) as a group-level effect. Examples of studies with typological samples that include varying intercepts for area in their regression models are Bentz et al. (2015), Bisang et al. (2020), Cysouw et al. (2012a), Lupyan and Dale (2010), Sinnemäki (2020), and Sinnemäki and Di Garbo (2018). Varying slopes by area are less commonly used, Jaeger et al. (2011) and Bentz and Winter (2013) being two notable exceptions.

Jaeger et al. (2011) propose another, innovative idea to include contact effects in a regression model by taking into account what neighboring languages do, using an aggregate measure of the dependent variable (in their case, phonological diversity) of neighboring languages as a predictor. To do so, they first calculate the distances between all pairs of languages in the sample. Then, for each language *L* in their sample, they calculate the aggregated neighbor phonological diversity by giving weights to each language of the sample according to its distance to language *L*. This results in languages close to *L* being strongly reflected in the additional neighbor predictor, while languages that are spoken at an increasingly larger distance from *L* are decreasingly represented by the neighbor predictor. As far as we know, this is the most convincing approach to modeling contact effects as they are theoretically assumed to arise, namely locally by neighboring languages in contact and by diffusion across larger areas over time, and not by large geographic areas as such. Nevertheless, we are not aware of another typological study that has implemented this approach.

One potential issue with Jaeger et al.'s approach is that the effect of distance is viewed as a constant across areas. It has long been known that linguistic diversity is not evenly distributed across the world (e.g. Hua et al. 2019; Nettle 1999; Nichols

1992), which means that languages can cover areas of very different sizes and have neighbors at very different distances. Also Dryer (2018), referring to Cysouw et al. (2012a), remarks that two languages spoken in areas like Siberia with a distance of 100 km may still share properties due to contact, while languages that are spoken 100 km apart in New Guinea are less likely to have been in contact.[23]

As a solution, Cysouw et al. (2012b) propose to control for contact bias by taking into account the number of languages spoken between a given pair of languages. This is implemented by Dryer (2018) as only including languages in the sample if they are not in the same genus and if there are at least 10 other languages spoken between them. The latter condition is operationalized as follows: "a language X is said to be between a language Y and a language Z if the distance between X and Y and the distance between X and Z are both less than the distance between Y and Z" (Dryer 2018: 803). This is a simple but elegant solution to control for language contact in the sampling process. As a modeling solution, it could be implemented as a weighted neighbor predictor using the approach proposed in Jaeger et al. (2011), replacing raw distances. Building on these ideas, the following section presents our proposal for modeling areal effects between languages within macroareas.

4.3.2 Our approach: Gaussian Process

To control for areal effects, we included a Gaussian Process (GP) term into our model with longitude and latitude data for each language.[24] A GP (Rasmussen 2003; Williams and Rasmussen 2006) is a relatively recent type of non-parametric method developed to handle non-linear dependencies in Bayesian regression models. The basic idea is that a GP can capture dependencies between all datapoints, and it does so to a different extent. In a GP, two close observations have a strong influence over each other, while two observations that are further apart will have little influence over each other.[25] An important feature of GPs is that they can induce the amount of smoothing automatically from the data, meaning that it

23 Rijkhoff et al. (1993: 174–175) make a similar remark about differences in distances between languages depending on the population density of the region.

24 For simplicity, we use Euclidean distances between languages. This is, of course, a simplification. It would be fairly easy to use other distance metrics with a GP, e.g. geodesic distances or walking distances (Wichmann and Hammarström 2020).

25 The readers might be more familiar with splines in generalized additive models, which have been used in e.g. dialectology to model areal effects (Wieling et al. 2011). Splines are effectively a special case of Gaussian Process (Kimeldorf and Wahba 1970). See also Baayen and Linke (to appear) for an introduction to generalized additive models in linguistics.

allows for the effect of distance to vary across areas rather than assuming that distances affects languages in the same way everywhere.

To illustrate the idea of a GP, Figure 6 displays a simple toy example of a non-linear dependency in a small dataset. Here, the data-generating model is $y \sim \sin(x) + \epsilon$, where ϵ is normally distributed noise ($\mu = 0$ and $\sigma = 0.2$). The red line represents the real data-generating process, while the dark blue line corresponds to the fitted GP and the shaded area is the 95% uncertainty interval of the model.

As can be seen in Figure 6, neighboring datapoints influence the estimates for other close values, e.g. relatively high y-value around x values between 8 and 9 lead to a smooth adjustment to higher estimates of y in that area. Another important property of a GP is that it has narrower uncertainty intervals in regions with more data and wider ones in regions with less or without data.

More formally, the model predicting y from GP(x) can be expressed as $y \sim N(m(x), K(x|\theta))$, meaning y follows a multivariate normal distribution, with mean $m(x)$ (where m is a mean function) and variance $K(x|\theta)$. The function K produces a covariance matrix on x given a set of parameters θ.[26] In our case, we use the

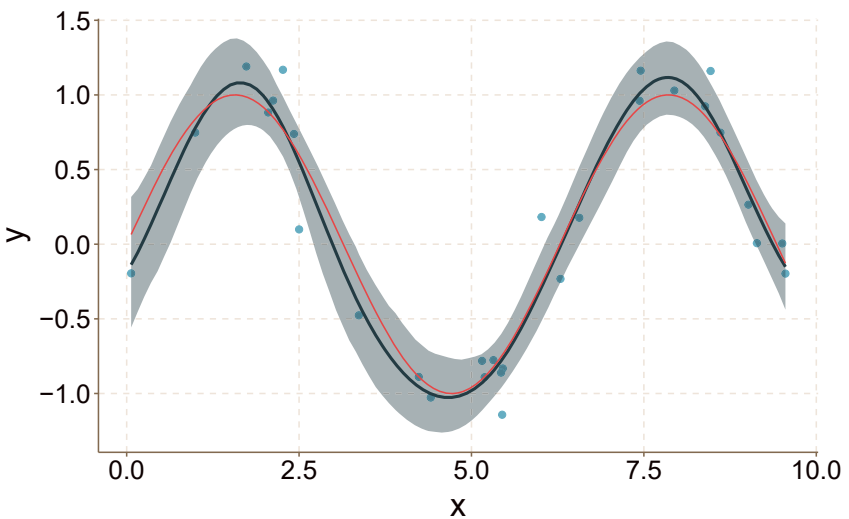

Figure 6: Example of a Gaussian Process fitted to a sinusoidal function.

26 Note that when using a non-Gaussian model we have to estimate a latent variable with the GP. This is the case in our study.

Exponential Quadratic Kernel to produce the covariance between all observations. The form of the Exponential Quadratic Kernel is: $\alpha^2 \exp\left(-\frac{0.5}{2\rho} \sum_{d=1}^{D} (x_{i,d} - x_{j,d})^2 \right)$, where i and j are indices for the observations, D the number of dimensions, ρ and α are parameters which control the strength of interactions and the rate of the decay of the interactions between observations. Figure 7 shows the estimated correlations for the example in 6; we transformed the covariance matrix to a correlation matrix to have the estimates in a scale from 0 to 1. The influence that each observation has on every other observation is shown with light blue lines, where the thickness of the line reflects the strength of the influence. The distance is solely based on the x-axis in Figure 7. As can be seen, points that are closer to each other have a stronger influence on each other (i.e. the model considers them to be similar) than points which are further apart. Points which are too far away from each other have effectively no influence on each other and thus, their correlation is not shown in the plot in Figure 7.

Since we aim at controlling for areal effects which come in two dimensions, we added a two-dimensional GP to our model. The two-dimensional GP is conceptually similar to the toy example from Figure 6. It acts as a surface which can capture dependencies across latitude and longitude of the languages in our sample.[27] We will assume that there has been relatively little contact across macroareas (for an example of potential contact across several macroareas along the Pacific Rim see Bickel and Nichols 2006), thus, we added independent GPs for

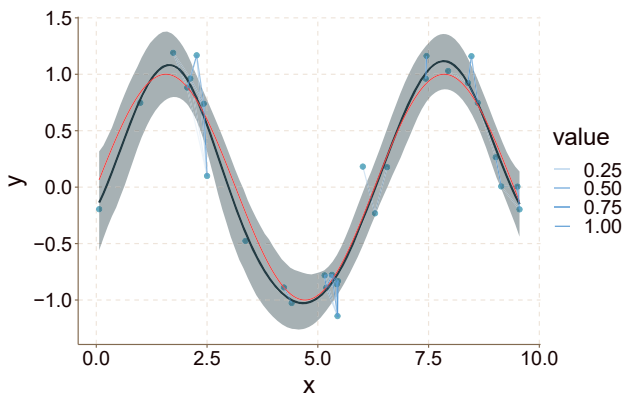

Figure 7: Example of a Gaussian Process fitted to a sinusoidal function, including the covariance matrix.

27 We used the latitude and longitude information provided in Glottolog.

each macroarea.[28] The results of the areal effects in our model are discussed in Section 5.4.

5 Results

5.1 Model definition

We model affix position as an ordinal response using a cumulative (i.e. ordinal) logit model. Using an ordinal model means that we assume that the response categories in our dataset are ordered with respect to each other. Our predictors are: (i) the verb-object order of the language, (ii) the phylogenetic term, and (iii) a two-dimensional GP on the (centered) longitude and (centered) latitude of the language by macroarea.[29,30] We added a Gaussian Process for each macroarea because we assume that there is little to no contact between macroareas, and that areal effects are confined to languages within each macroarea. While this assumption might be too strong in some macroarea borders, the intuition that South American languages and African languages have had effectively no contact seems adequate and is captured by this implementation. We fitted the model using Stan (Carpenter et al. 2017) with the brms interface (Bürkner 2017, 2018)[31] in R (R Core Team 2020).[32]

28 In principle, one could use a single GP for the whole dataset, and it should work similarly to our approach. However, adding a single GP makes predictions more difficult for the model, as it has to estimate on its own that there has been effectively no contact between South American languages and African languages, or between Australian and Eurasian languages. The decision to include information on macroareas in the GP can be viewed as an additional prior on the model. In practice, we also tried to use a single GP. Without any clear predictive differences, the models using a single GP took about three times as long to fit as the models using multiple GPs. Determining whether a single GP or multiple GPs is better to account for contact and areal effects would go far beyond the scope of this study.

29 The brms formula is affixation ~1 + verb-object order + gp(longitude, latitude, by = macroarea) + (1|gr(micro family, cov = phylogeny)).

30 We used mildly informative priors for most parameters of the model. For the phylogenetic term we set a strong prior (normal $\mu = 1$, $\sigma = 0.1$) on the standard deviation to prevent the model from overfitting. A weaker prior leads the model to overfit for isolates because these are free to vary independently of all other languages. The full model specification can be found in the Supplementary Materials. For the GP we used the default priors provided by brms for the length scale, and a mildly informative prior (normal $\mu = 0$, $\sigma = 1$) for the standard deviation.

31 Version 2.15.

32 Stan is a probabilistic programming language for Bayesian inference. While we work within a Bayesian framework for the models in this paper, the approach presented here is compatible with a frequentist framework as well.

We use visual inspection of conditional effect plots for the evaluation of the models. Those plots show how the outcome changes when one of the predictors changes while the others remain constant. We use conditional effect plots rather than trying to interpret the coefficients of the estimated parameters directly, since visual inspection of the effects of predictors on the outcome variable is much more straightforward. This is especially important with models that are very complex or contain non-linear effects (as with the GP terms).[33]

It is important to emphasize that we are not doing null hypothesis significance testing. We are therefore less concerned with the question of whether two estimates are significantly different from each other in the statistical sense. That is, we do not assess the probability of the data given the null hypothesis. Instead, we estimate the probabilities that two estimates are different from each other directly. Partial overlap of the error bars (or the individual draws) does not mean that two estimates are not significantly different, but rather, that the difference is not as certain (the exact probability of this difference can also be estimated).

5.2 Effects of verb-object order

First we examine the effect of verb-object order on affix position across languages. The conditional effects are shown in Figure 8. As mentioned above, the conditional effects plot shows how the effect of verb-object order on the probability of each values of affix position changes, holding all other predictors constant. Using Bayesian inference, the error bars in Figure 8 (and in the following plots in Section 5) represent the posterior 50% (thick bars) and 95% (thin bars) uncertainty intervals. Importantly, an uncertainty interval is different from a confidence interval. The uncertainty interval is the region in which 50% or 95% of the values of the posterior lie. The dots represent the mean value of the posterior, that is, our best guess as to what the real value of the parameter (in this case, the predicted response probability) might be.[34]

As can be seen from the similar mean estimates and also from the largely overlapping uncertainty intervals in Figure 8, the model estimates a rather small mean effect of verb-object order for each value of affix position. This very weak effect seems to come from VO-languages which avoid being strongly suffixing. Given the wide uncertainty intervals, however, we have to conclude that when including phylogenetic and areal information in the model, the signal from effects of verb-

33 See also Gabry et al. (2019) for the importance of visualization for interpreting model results.
34 For reasons of space, in this paper we will focus only on the mean values and uncertainty intervals instead of doing a full posterior analysis. For a description on alternative ways of exploring the posterior, see Gelman and Loken (2013) and Gabry et al. (2019).

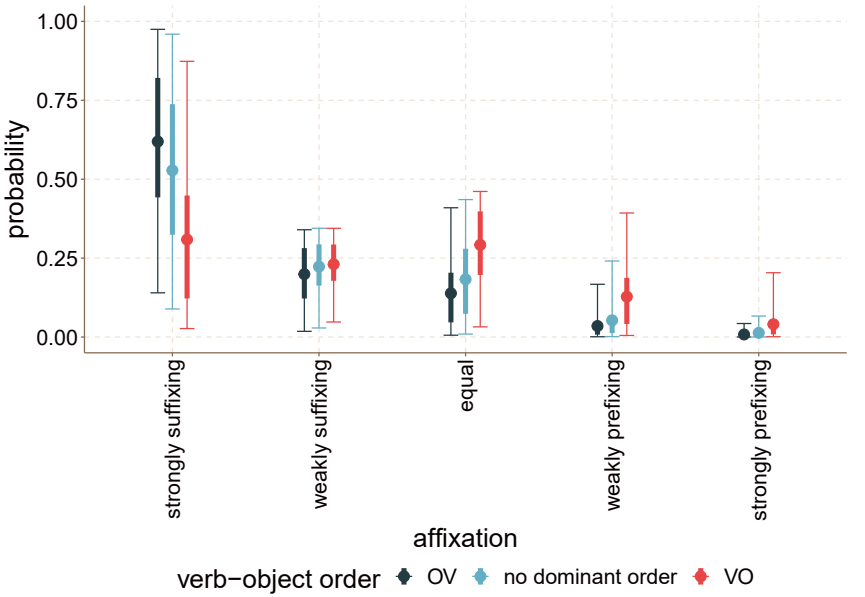

Figure 8: Conditional effects of verb-object order on affix position.

object order on affix position is not strong enough for a substantial degree of certainty. The likely explanation for why we see such wide uncertainty intervals in the conditional effects is that the apparent correlation between word order and affix position is mediated by both genealogical and areal effects. That is, the genealogical and areal effects are a good explanation of the distribution of both word order and affix position.[35] We therefore conclude that, according to our model, there does not seem to be any effect of verb-object order on the position of the affix.

5.3 Phylogenetic effects

Figure 9 shows the group-level effect estimates for the Slavic, Germanic, Romance, and Athabaskan micro-families (50% uncertainty intervals are shown as thick lines, and 90% uncertainty intervals are shown as thin lines with whiskers). We

35 One anonymous reviewer suggested that the very wide uncertainty intervals were likely due to the sampler having difficulties in identifying the parameters of the model. This is unlikely. All test statistics (Effective Sample Size, R-hat, etc.) were within normal ranges and we did not see any errors in the sampling (e.g. divergences). A simple illustration of why our explanation is likely can be found in the Supplementary Materials.

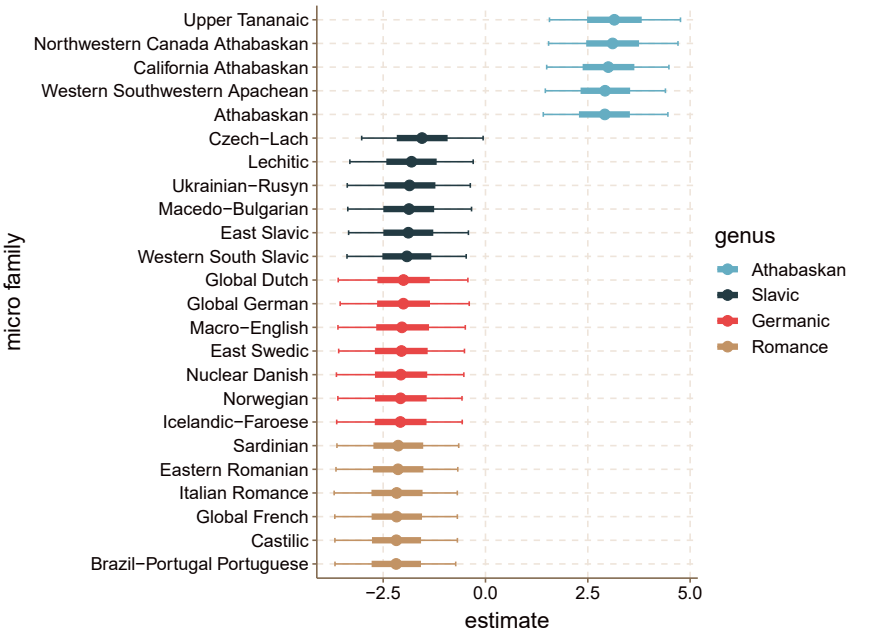

Figure 9: Phylogenetic effects for Slavic, Germanic, Romance, and Athabaskan languages.

selected these languages for the purposes of illustration. More specifically, it shows the intercepts of each micro-family, which are forced to covary according to the phylogenetic structure between and within families. The exact values shown in Figure 9 itself work in relation to the rest of the model, where smaller values correspond to a suffixation preference, and larger values stand for prefixation.

Thus, we see that the estimates for micro-families which are genetically very close are very similar to each other as well. For instance, all Romance languages in the sample are estimated to have very similar intercepts for affix position. Also for the Slavic micro-families, Figure 9 shows that all intercept estimates are very similar to each other, even though Czech is classified as weakly suffixing, with all other Slavic languages in the sample annotated as strongly suffixing. We can see that the estimate for Czech-Lech therefore also differs slightly more from the ones of the other Slavic micro-families, but the phylogenetic term keeps the estimated variation between closely related languages low. It is important to note that while in this instance the structure of the phylogenetic effects seems to resemble genera, this is not necessarily the case. In situations of strong variation within genera the effects of each language will vary more from the mean of related languages.

In contrast to Romance, German, and Slavic, Figure 9 shows much higher values for the estimates for Athabaskan micro-families. This is partly due to the fact that Athabaskan languages are simply classified as having more prefixation, but the intercepts of Athabaskan are also allowed to be very far from the intercepts of Indo-European languages because the micro-families are not related to each other in the phylogenetic tree. The intercepts for the Athabaskan genus are very close to each other even though we find variation in the annotated affix position, with Western Apache (Western Southwestern Apachean), as equally prefixing and suffixing, with Sarsi (Athabaskan), Hupa-Chilula (California Athabaskan) as weakly prefixing, and with Navajo (Western Southwestern Apachean), Chipeywan (Northwestern Canada Athabaskan), and Tanacross (Upper Tananic) as strongly prefixing. These differences are reflected in slightly higher estimates for the latter two micro-family intercepts. Again, the phylogenetic term causes the intercept estimates of closely related languages to be very similar; it allows the ones of less closely related languages to vary somewhat more, and the ones of unrelated languages to vary freely.

This allows for a more fine-grained control over the genealogical effects than simply using family (or genus), but at the same time, it prevents overfitting the model. If we were to add varying intercepts for each micro-family without the phylogenetic term, this would cause the model to have almost perfect knowledge about the data at hand, but it would not be able to generalize to new data.

5.4 Areal effects

As mentioned above, we use a Gaussian Process as a way to control for non-linear, areal or contact effects, as introduced in Section 4.3.2. Visualizing the joint conditional effect of longitude and latitude, i.e. the effect of the GP, can be done by a contour plot. Figures 10–15 are such contour plots, showing the surface which captures large-scale areal effects, with the languages in our sample marked as red dots. Figures 11–15 show the GP effects of longitude and latitude for the six macroareas of Africa, Eurasia, Papunesia, Australia, North America, and South America.

The plots should be read like elevation maps, with hills (red) and valleys (dark blue).[36] The valleys are associated with suffixing languages (closer to 1), the hills are associated with prefixing languages (closer to 5). Note that the model makes

36 Note again that the effects displayed in the maps are not absolute predictions of the model, but rather how the prediction of the model varies when we vary latitude and longitude and set the other covariates (i.e. verb-object order) to a fixed value (except the phylogenetic term).

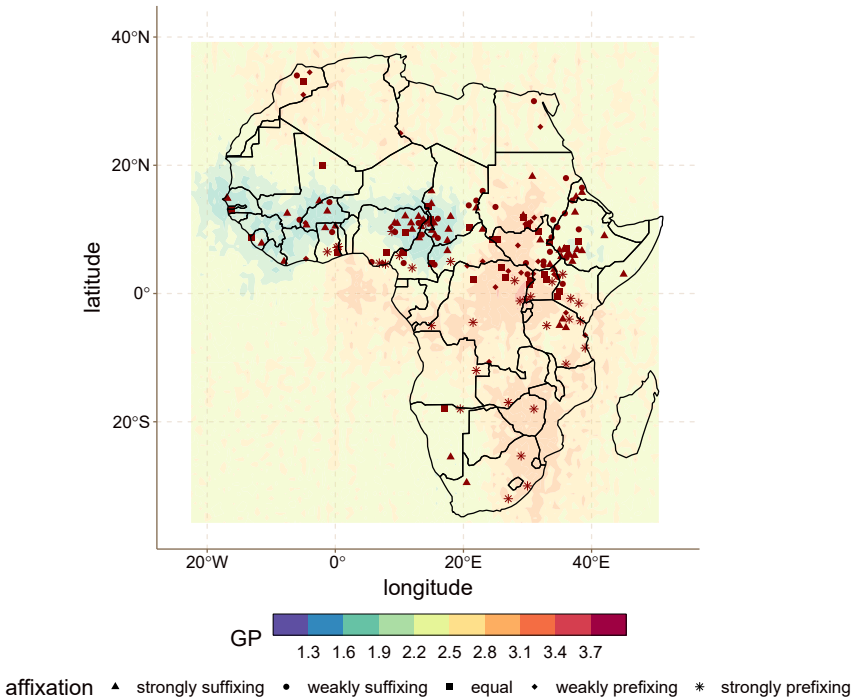

Figure 10: Geographic effects in Africa.

estimations for all geographical areas, including for regions with no languages in the sample (e.g. for the center of the Pacific Ocean). Such estimates correspond to regions with high uncertainty (it is not possible to properly display the uncertainty on the maps in this section; however, see the Supplementary Materials).

Looking at single macroareas, Figure 10 confirms the well-known pattern in Africa in that East Africa and Southern Africa have a stronger preference for prefixation in comparison to West Africa, where we find some clusters of strong suffixation. This is hardly surprising given that the dataset contains 22 Bantu languages which are classified as strongly prefixing and which are spoken predominantly in Central, East, and Southern Africa. Figure 10 shows a Western-most suffixing area from Senegal to Ghana. The next suffixing hotspot is located between Niger, Chad, and Cameroon, and is likely due to a number of Chadic languages in the sample. The third strongly suffixing area in Africa is located mainly between Eritrea, Ethiopia, Sudan, and South Sudan, in part because of several Cushitic and Semitic suffixing languages in the sample.

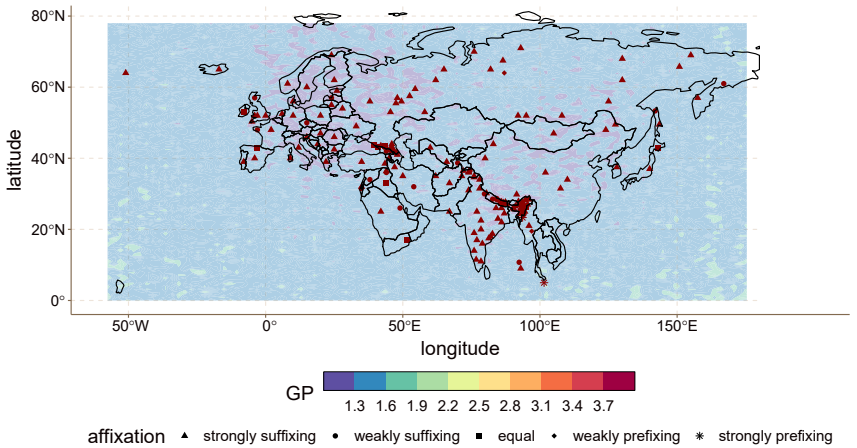

Figure 11: Geographic effects in Eurasia.

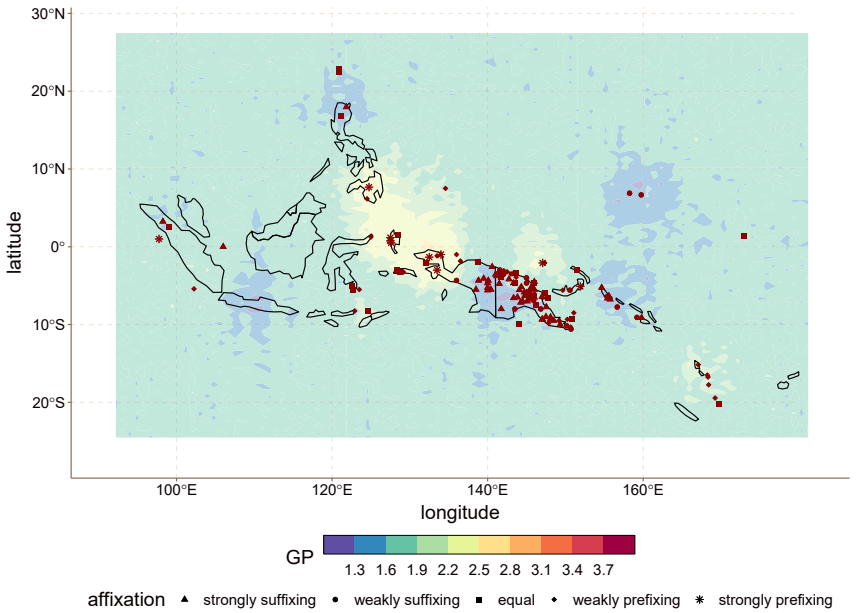

Figure 12: Geographic effects in Papunesia.

For Eurasia, as can be seen in Figure 11, the areal effect is almost non-existent. This is due to the low degree of variation in affix position within Eurasia, which is predominantly suffixing as a whole. Out of 181 languages in this macroarea, only 4

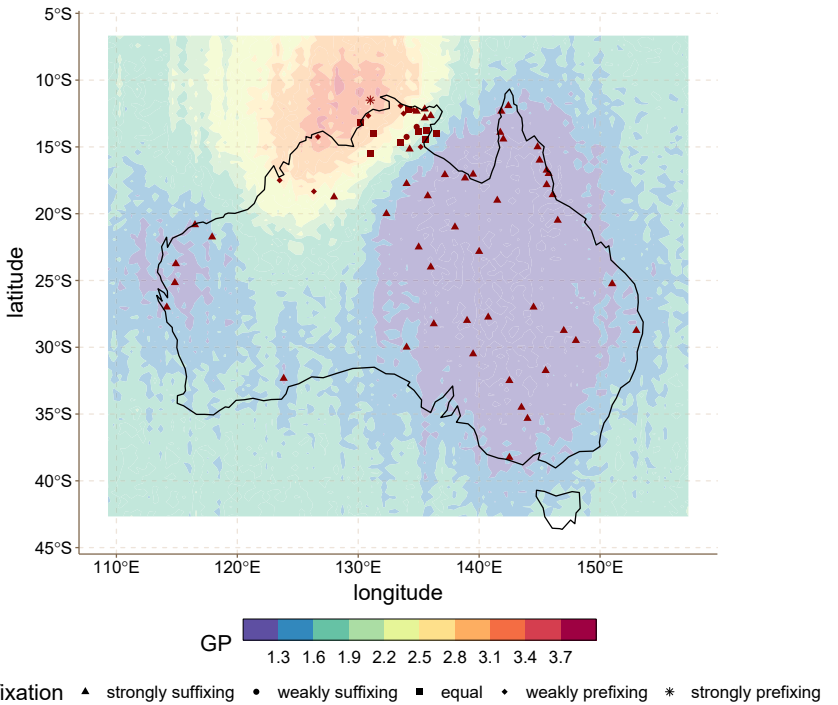

Figure 13: Geographic effects in Australia.

(Tedim Chin, Ket, Temiar, Bwe Karen) are classified as weakly prefixing or strongly prefixing, while all others are either coded as being equally prefixing and suffixing or having a suffixation preference.

In Figure 12, we see a relatively strong effect towards prefixation around the Maluku islands in the West and the Bird's Head Peninsula in Papua in the East. This is due to a rather high number of weakly or strongly prefixing languages from different top-level families: West Makian, Sahu, and Tidore (North Halmahera) on the Maluku islands; Maybrat-Karon (Maybrat-Karon), Irarutu, Biak, and Ambai (Austronesian), Hatam (Hatam-Mansim), and Meyah (East Bird's Head) on the Bird's Head Peninsula in West New Guinea. At the same time, there are no languages from the sample in those two areas which are classified as suffixing languages. In the Central and Eastern parts of New Guinea, on the other hand, the sample includes a large number of Nuclear Trans New Guinea languages which are mostly all strongly suffixing.

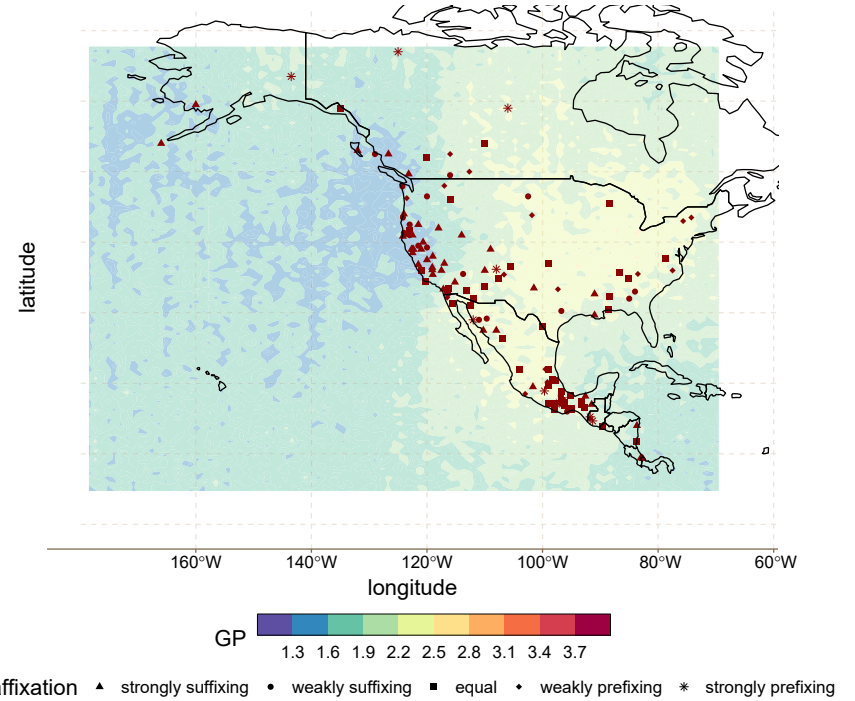

Figure 14: Geographic effects in North America.

For Australia, Figure 13 shows a very strong difference between the northern parts of the Northern Territory and Western Australia, and the remaining parts of the continent. The northern part has a fairly strong preference for prefixation, with a clear tendency toward suffixation in the rest of the continent. This is also expected, given the well-known division between Pama-Nyungan and non-Pama-Nyungan languages, the latter of which are spoken in the northern parts of Australia with rich verbal prefixation. Pama-Nyungan languages, on the other hand, cover most of the continent and exhibit little prefixation, reflected in the classification of all 43 Pama-Nyungan languages in the sample as strongly suffixing. Nine of the non-Pama-Nyungan languages in the sample are classified as having a prefixing preference, and the ten other non-Pama-Nyungan languages are coded as equally prefixing and suffixing. Note that the phylogenetic relations between Pama-Nyungan languages and non-Pama-Nyungan languages on the one

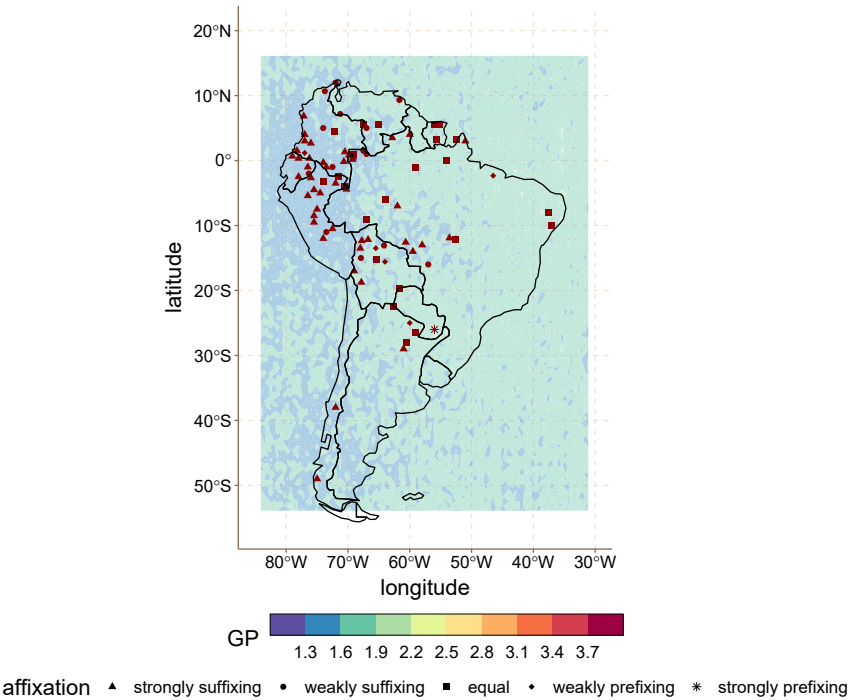

Figure 15: Geographic effects in South America.

hand, as well as within non-Pama-Nyungan is still a topic of debate (see Koch 2014 for an overview). Also, the crude division between Pama-Nyungan and non-Pama-Nyungan based on affix position draws a very simplified picture. For instance, Kayardild is a strongly suffixing Tangkic language (also in the sample), whereas Yanyuwa, spoken in the prefixing area, is a prefixing Pama-Nyungan language (Koch 2014: 64). This underlines the importance of areal effects, contact, and diffusion and the need to control for these as well.

Figure 14 shows the areal effects for North America including Central America. We can see a split between the West Coast in North America on the one hand and the East Coast and Central America on the other. The West Coast, especially California and Oregon, has a clear preference for suffixation. A large number of Penutian languages in the sample are suffixing languages spoken in this area, as well as 12 Uto-Aztecan (mostly Numic) suffixing languages. The Uto-Aztecan languages spoken further to the south in Mexico, on the other hand, are mostly prefixing and contribute to the prefixing preference that we see in the central, eastern, and southern parts of North America. Other families that likely contribute

to the slight prefixing preference in the center of North America are Na-Dene and Maya, with eight equally prefixing and suffixing Oto-Manguean and Uto-Aztecan languages spoken in the central and southern areas of North America. In the east, the sample includes four prefixing Iroquoian languages.

South America does not have strong areal effects, as is shown in Figure 15. The slightly stronger trend towards suffixation in the West is likely due to the large number of strongly suffixing languages in Colombia, Ecuador, and Peru. These include a number of Arawakan, Cariban, and Tucanoan languages, amongst other languages from many other families. Of the six languages with a prefixation preference, three belong to the Tupian family. Their distribution over the continent, however, does not appear to show any strong areal trend.

Overall, these results show that areal patterns clearly play an important role (cf. Section 6.3), and a Gaussian Process is able to detect and to control for them. The "intuitive" areas which are strongly associated with either suffixation or prefixation are clearly visible on the maps in Figures 10–15, and the model is flexible enough to detect patterns at short distances (e.g. in Papunesia) or at long distances (e.g. in Africa).

5.5 Model evaluation and performance

Model evaluation is important, as it tells us how reliable the results are for interpretation and analysis. The concept behind a posterior predictive check is summarized by Gelman and Hill (2007: 513) as follows:

> Monitoring the quality of a statistical model implies the detection of systematic differences between the model and observed data. Posterior predictive checks set this up by generating replicated datasets from the predictive distribution of the fitted model; these replicated datasets are then compared to the observed dataset with respect to any features of interest.

Thus, to assess the performance of our main model, we first perform a visual posterior predictive check following Gabry et al. (2019). This means that we use the fitted model to generate data and compare the distribution of predicted affixation patterns with the observed one in the dataset. Ideally, the distribution of posterior samples (the predicted data) should closely match the observed distribution in the dataset. This is an important indication for how useful and appropriate the specified model is to represent our data.

Figure 16 shows the distribution of posterior samples from the model, i.e. the predicted data (light blue), and the observed data (red). It reveals that the distribution generated by the model closely matches the distribution of the observed

data. Hence, we can assume that the model is correctly specified, and that an ordinal model is appropriate for our data.

Additionally, we performed approximate leave-one-out cross-validation (LOO-CV) using the method described by Vehtari et al. (2017). The idea of the LOO-CV is to fit the model on the data leaving out a single observation and to predict this observation. This process is repeated for all observations, and it serves as a way to assess how well the model can predict unseen data (and if, for instance, the model is overfitted or depends too much on single datapoints). The method proposed by Vehtari et al. (2017) which we applied here is a more resource-friendly approximation of the LOO-CV, but its idea and purpose remain the same.

With approximate LOO-CV, we use three different metrics to evaluate the performance of the models: accuracy, kappa score, and root mean square error (rmse). The reason for combining metrics from discrete classification (accuracy and kappa score) and regression (rmse) is that ordinal regression models are somewhat of a hybrid between both (see Gaudette and Japkowicz 2009: for a discussion of evaluation metrics for ordinal regression).

The accuracy measure simply corresponds to the number of correct responses divided by the total number of observations. The kappa score is similar to accuracy; it ranges from 0 to 1 and it captures how much better than random chance the

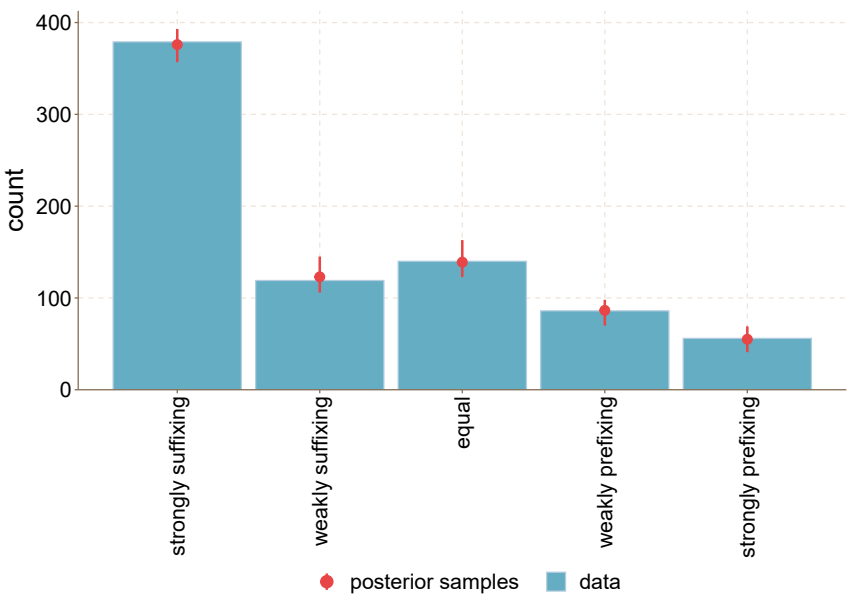

Figure 16: Posterior predictive check.

model is performing. This means that it does not only take into account the number of correctly classified responses, but it also takes into account the degree of variation that the response variable has. If 95% of all observations were strongly suffixing, simply classifying all observations as strongly suffixing would lead to an accuracy of 0.95. In such a scenario, guessing at chance can nevertheless lead to a high accuracy because of the low degree of variation in the response. Obviously, with more variation, classifying the responses correctly at chance becomes harder. Thus, the kappa metric indicates how much better the model classifies the responses compared to guessing at random chance, given the variation in the response.

The metric rmse is calculated as $\sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2}$, with smaller values indicating a better model fit than larger values. In contrast to the accuracy and kappa measures, rmse does not only measure how well the model performs on exact predictions, but it also captures how far off incorrect predictions are from the real value in general. This is an important measure for ordinal data with inherently ordered categories. Intuitively, if the model predicts a strongly suffixing observation to be weakly suffixing, this prediction would be less wrong than if the model classified the observation as strongly prefixing.

Table 2 shows a confusion matrix and the three evaluation metrics of the approximate LOO-CV. The columns of the confusion matrix display the number of observations for each affixation value in the data. In the rows, we see the predictions made by the main model. We find the observations which are correctly classified by the model in the diagonal of Table 2.

Table 2: Confusion matrix and evaluation metrics for the main model.

	Reference				
Prediction	**Strongly suffixing**	**Weakly suffixing**	**Equal**	**Weakly prefixing**	**Strongly prefixing**
Strongly suffixing	**230**	24	7	1	0
Weakly suffixing	124	**66**	47	16	5
Equal	23	26	**70**	47	18
Weakly prefixing	2	3	16	**20**	28
Strongly prefixing	0	0	0	2	**6**
Accuracy			0.5		
Kappa			0.32		
Rmse			0.88		

The bold values correspond to the number of correctly predicted observations.

Overall, the model performs reasonably well. Although an accuracy of 0.5 and a kappa value of 0.32 do not seem very high, the confusion matrix in Table 2 shows that most errors only occur between neighboring values. For instance, the model correctly classifies 230 observations as strongly suffixing, and wrongly predicts 124 observations to be weakly suffixing instead of strongly suffixing. It also misclassifies 23 strongly suffixing observations as equally prefixing and suffixing, and two as being weakly prefixing. Crucially, the model does not predict any strongly suffixing observation to be strongly prefixing. We also see that the model struggles most with the classification of strongly prefixing observations, which can be explained by their generally low frequency in the data. Despite this, the majority of misclassifications are limited to adjacent affixation values. The rmse value of 0.88 is not very informative as such but can be used for model comparison, as we will show in Section 6.

6 Model comparison and robustness

In Sections 4 and 5, we discussed in detail our proposal of a model which controls for genealogical and areal biases. However, there are many alternative models. While we cannot explore all possible model specifications, we will comment on a few alternative models, focusing on two in particular: a hierarchical model (Section 6.1) similar to the one proposed by Jaeger et al. (2011) and a no-controls model (Section 6.2). These two model specifications are relevant, since they are fairly standard alternatives to the main model proposed in this paper.

6.1 Hierarchical model

The hierarchical model that we fitted had the formula of: affixation ~VO-order + (1|family) + (1|macroarea).[37] Thus, its predictors include the population-level effect of interest, i.e. verb-object order, and varying intercepts by family and macroarea (group-level effects).

Figure 17 shows the conditional effects of verb-object order for the hierarchical model. The most important finding here is that with our main model being much less certain about the effect of verb-object order on the position of the affix, the hierarchical model is much more certain. The hierarchical model also estimates a much clearer difference in effect between OV and VO-orders, with OV-languages

37 We also explored the possibility of using genus instead of family but this led to difficulties fitting the model and no better performance.

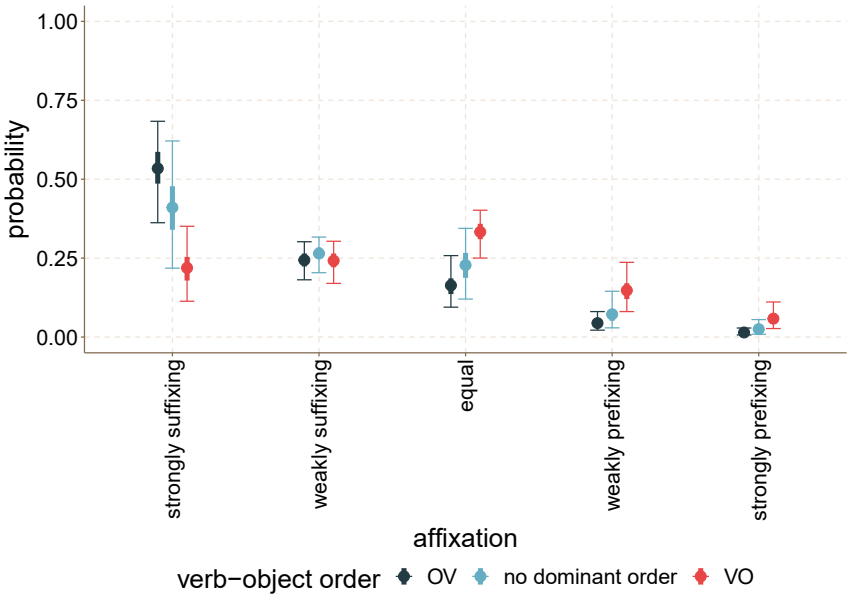

Figure 17: Conditional effects of verb-object order on affixation preference for the hierarchical model.

showing a marked preference for strongly suffixing languages when compared to VO-languages. The relation is then reversed for equally prefixing and suffixing, weakly prefixing and strongly prefixing.

As mentioned before, our aim is not to confirm or to reject a hypothesis, but rather to estimate the degree of uncertainty of the effects. If we were to interpret the results of the hierarchical model, we might be led to conclude that an effect of verb-object order on affix position is very likely. In contrast, our main model including phylogenetic and areal controls as described in Section 4 is much more conservative in its estimates and suggests that an effect of verb-object order on affix position is very unlikely and that the differences are likely due to random variation.

In order to assess how well the hierarchical model captures the data, Table 3 shows the model performance in the approximate LOO-CV.[38] In terms of rmse (0.97), accuracy (0.47), and kappa (0.26) values, the performance of the hierarchical model is worse than of the main model (rmse = 0.88, accuracy = 0.5, and kappa = 0.32). Nevertheless, looking at the confusion matrix in Table 3, the

38 We do not use p-values, R^2, or their Bayesian equivalents in this paper because they have been shown not to be useful for model comparison or feature selection (Faraway 2006; Gelman and Loken 2013, 2014).

hierarchical model still manages to make reasonable predictions for most categories. However, Table 3 also shows that the model is very biased against strongly prefixing languages, with zero predictions of this value.

This section showed that the traditional approach using group-level effects of family and macroarea can control for some parts of bias in the estimate and perform adequately well, but it produces biased estimates for the effect of verb-object order, and its predictions are poorer than that of our main model.

6.2 No-controls model

The model without any statistical controls for bias may reflect a situation in which all bias control is assumed to be part of the sampling process or to be unnecessary, given the large sample size of 780 languages. The no-controls model thus only includes the linguistic predictor of interest, namely verb-object order.

Figure 18 shows the conditional effects for verb-object order for the no-controls model. Compared to both the main and hierarchical model, the no-controls model estimates a much larger effect of verb-object order on affix position. The model estimates that VO-languages are much less likely to be strongly suffixing than OV-languages and languages without dominant order. The no-controls model also suggests that VO-languages are more likely to be weakly or strongly prefixing than OV-languages and languages without a dominant order.

This model has much smaller uncertainty intervals as well, meaning that it is much more certain about the estimates and their differences.

Table 3: Confusion matrix and evaluation metrics for the model with hierarchical control for family and macroarea bias.

	Reference				
Prediction	**Strongly suffixing**	**Weakly suffixing**	**Equal**	**Weakly prefixing**	**Strongly prefixing**
Strongly suffixing	**247**	36	6	0	2
Weakly suffixing	93	**40**	56	35	4
Equal	28	37	**64**	36	12
Weakly prefixing	11	6	14	**15**	38
Strongly prefixing	0	0	0	0	**0**
Accuracy	0.47				
Kappa	0.26				
Rmse	0.97				

The bold values correspond to the number of correctly predicted observations.

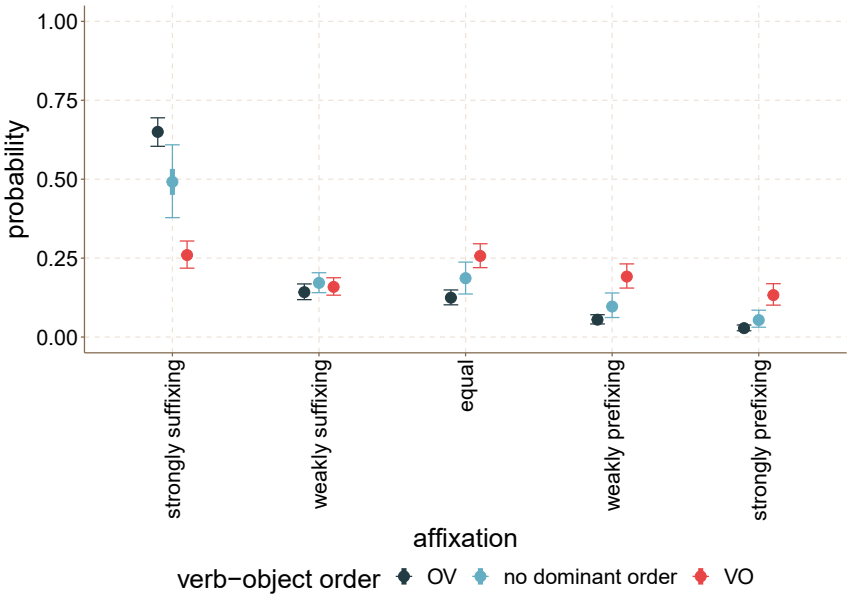

Figure 18: Conditional effects of verb-object order on affix position for no-controls model.

Table 4 shows the model performance in the approximate LOO-CV. This model clearly under-performs and fails to correctly classify most observations in their affix position. Its accuracy is only 0.2, its kappa score is close to 0 and its rmse of 1.2 is much higher than for the main (0.88) or the hierarchical (0.97) model.

Table 4: Confusion matrix and evaluation metrics for the no-controls model.

	Reference				
Prediction	**Strongly suffixing**	**Weakly suffixing**	**Equal**	**Weakly prefixing**	**Strongly prefixing**
Strongly suffixing	**0**	0	0	0	0
Weakly suffixing	286	**77**	64	29	5
Equal	93	42	**76**	57	51
Weakly prefixing	0	0	0	**0**	0
Strongly prefixing	0	0	0	0	**0**
Accuracy	0.2				
Kappa	0.04				
Rmse	1.2				

The bold values correspond to the number of correctly predicted observations.

This is an important result because it tells us that using the WALS sample (or any other large convenience or variety sample, for that matter) as such leads to a strong bias in the estimated effect.[39] It is thus not sufficient to control for genealogical and areal bias by simply using a large variety sample of languages. Such a sample can be used, though, together with at least some form of statistical controls which help to improve the models and the estimates.

6.3 Other model specifications

Our main model presented in Section 4 has three components: a phylogenetic term, an areal effect, and the main effect of interest for verb-object order. In the previous sections, we discussed how the phylogenetic and areal terms help us to control for bias in the estimate of the verb-object order effect. However, their relative contribution to the model has not yet been assessed.

To explore to what extent each component contributes to the model, we fitted multiple models leaving out one of the effects at a time. Those models can be compared using ELPD (expected log predictive density), which is calculated using approximate LOO-CV (Vehtari et al. 2017). ELPD values are difficult to interpret as such, but a model with a larger ELPD value is expected to have better predictive performance than a model with a lower ELPD value.

Table 5 shows the model comparison using ELPD. The column ELPD diff shows the difference in ELPD between the best model (1) and each of the 7 other comparison models. The column SE diff shows the standard error of the estimate. Since the ELPD is normally distributed, it is often assumed that we can be certain that a difference in ELPD is real if the difference is greater than two times the standard error.[40] The ELPD comparison in Table 5 shows that the main model has the best expected predictive power. In other words, if we were to try to predict new data, we would expect the main model to make the most accurate predictions. The model without the main effect for verb-object order (2) is the second best model in the list, and, although its ELPD is lower than that of the main model, the standard error of the difference is relatively high. This result confirms that it is unlikely that there is a real effect from verb-object order on affix position.

The next model in the comparison which excludes the areal term (3) has clearly lower ELPD than the main model with an ELPD difference of 15.9. This shows that

39 This is not a point of criticism against the WALS database. On the contrary, we are fully aware that one of the aims was to include as many languages as possible, and that it was not purely designed to provide balanced samples (cf. Comrie et al. 2013).

40 However, the SE may be optimistic, which is why a difference of at least four times the SE is considered as more conservative (Vehtari et al. 2021). Using the more conservative threshold of four does not change much of the interpretation of the results in this section.

Table 5: ELPD differences between the main and seven alternative models.

Model			ELPD diff	SE diff
1	phylo + areal GP + verb-object	(main)	0.0	0.0
2	phylo + areal GP		−10.0	5.8
3	phylo + verb-object		−15.9	6.1
4	(1|family) + areal GP + verb-object		−16.1	7.2
5	(1|family) + (1|macroarea) + verb-object	(hierarchical)	−55.7	10.5
6	(1|family) + verb-object		−55.9	10.7
7	areal GP + verb-object		−72.9	10.7
8	verb-object	(no controls)	−221.1	14.5

removing areal controls has a negative impact on the expected predictive performance of the model. Thus, without areal information, the model can predict affix position less well for new languages. We see a similar decrease in ELPD (16.1) when we replace the phylogenetic term in the main model with a simple group-level effect for family (4). Again, this suggests that using a phylogenetic term is better than simply including family as a group-level effect.

A more dramatic drop in ELPD can be observed moving to the hierarchical model (5). That is, if we remove the areal GP and phylogenetic controls, replacing them with varying intercepts by family and macroarea, the model performs yet considerably worse with an ELPD difference to the main model of 55.7.[41] The next model with only varying intercepts by family as a control (6) performs similarly to the hierarchical model (5) with group-level intercepts for both family and macroarea. Thus, adding the group-level effect for macroarea as in the hierarchical model makes essentially no difference and it does not appear to be an effective areal control. Model 7 shows that completely removing all controls for family bias has a dramatic impact on model performance, which we will discuss below. The model without any controls (8) is clearly the worst performing model.

Disentangling the relative importance of areal and phylogenetic effects is difficult. The reason is that if we remove the areal GP component, a portion of the variance falls onto the phylogenetic term because closely related languages also tend to be spoken in geographic proximity to each other. Even though Table 5 suggests that one-level family and phylogenetic effects might be stronger than

41 We also tried different model specifications which performed equally as well as our main model but which we do not describe here to save space. Some of these were using a non-isotropic Gaussian Process and tensor products instead of the isotropic Gaussian Process, fitting a single GP instead of six independent ones, as well as different specifications for the phylogenetic term. We chose the model which was computationally simpler, and, in our opinion, more intuitive from a linguistic perspective.

areal effects, this is not completely clear because the phylogenetic term is flexible enough to cover part of the areal effects once we remove the areal term.

Model 7 shows that completely removing all controls for genealogical bias has a dramatic impact on model performance, the ELPD difference to the main model being 72.9. However, if we compare models 6 and 7 directly as in Table 6, we observe that removing areal effects from the main model (and replacing the phylogenetic term with the less flexible family effect) as in model 6 makes the model similarly worse for predicting the affix position in new languages as when including an areal GP control only as in model 7. What this effectively shows is that the areal controls are about as important as the family controls, at least for affixation position and this dataset.

To sum up, these results show (i) that the main model is the model with the best expected predictive performance; (ii) that despite being less optimal than the phylogenetic approach, using family controls as in models 4, 5, and 6 increases the performance of the model considerably over model 8 without any controls; that (iii) areal effects are at least as important as family effects; and (iv) that simply adding varying intercepts by macroarea to a model is not sufficient to control for areal biases.

6.4 Oversampling

One of the main claims of this paper is that statistical controls can cope well with biases in the sampling procedure. To test this claim, we simulate the effects of two strongly biased samples.

First, we simulate a scenario in which a single language family is overrepresented in the sample. To do so, we selected the following 10 languages and added each of them 10 times to the sample: Italian, Swedish, Dutch, Danish, Czech, Slovenian, Irish, Welsh, Tajik, and Central Kurdish. This produced a new dataset with 100 additional Indo-European datapoints, which is equivalent to overrepresenting this language family in the sample. This is a relevant test case especially given the historically grown and still prevalent bibliographical bias

Table 6: ELPD differences for models 6 and 7.

		ELPD diff	SE diff
6	(1|family) + verb-object	0.0	0.0
7	areal GP + verb-object	−16.7	15.4

towards over-representing Indo-European languages. If our phylogenetic control works as we assume, we expect the model fitted on the oversampled data to produce similar predictions to the model trained on the original dataset.

The second test case is a scenario of an oversampled macroarea. For this simulation, we added each South American language three times to the sample, which results in 74 additional languages. In both the Indo-European and South American oversampled datasets, we added a certain amount of jitter to the coordinates of all additional languages to avoid two languages being on top of each other.

We then fitted the following three models to these datasets: the main; the hierarchical; and the no-controls model. The conditional effects are shown in Figure 19 for the original sample (left column), the Indo-European biased sample (mid column), and the South American biased sample (right column). We combine these samples with our main model including controls for contact and areal bias (top row), the hierarchical model (mid row), and the no-controls model (bottom row).

For the main and the hierarchical model, Figure 19 shows that the estimated effects are very similar across datasets. With some fluctuation of the mean estimates between the original and the oversampled datasets, their uncertainty intervals cover most of the same regions. For the no-controls model, we find a minimal shift of the estimate of the effect of VO-order on the probability of strongly suffixing, but not much more.

Table 7 shows the ELPD comparison for the oversampled datasets for the main (1), the hierarchical (2), and the no-controls (3) models. As we saw before, the main model performs better than the hierarchical model and the no-controls model. Interestingly, in both cases of oversampling, the hierarchical model performs worse than it did with the original dataset (ELPD difference of 75.6 for South America and 120.4 for Indo European with oversampled data, and ELPD difference of 55.7 with original data), suggesting that the hierarchical model has more difficulties dealing with oversampled datasets than the main model.

The results of this section show that moderate oversampling of large samples is not a noticeable problem as long as statistical controls are used and as long as the remaining families or areas are not heavily under-represented in the sample. In the case of the no-controls model, moderate oversampling did seem to have a minor but noticeable impact on the estimates and on their uncertainty.

This does not mean, however, that any sample will produce the same results. We also tested the main model on smaller random sub-samples of 100 languages (not reported here for reasons of space), which is a common sample size for typological studies. Such sample sizes resulted in the strong instability of the model estimates. We can therefore say that oversampling a family or a region can statistically be controlled for, as long as one includes as many languages from different groups and areas as possible. Another insight emerging as a by-product

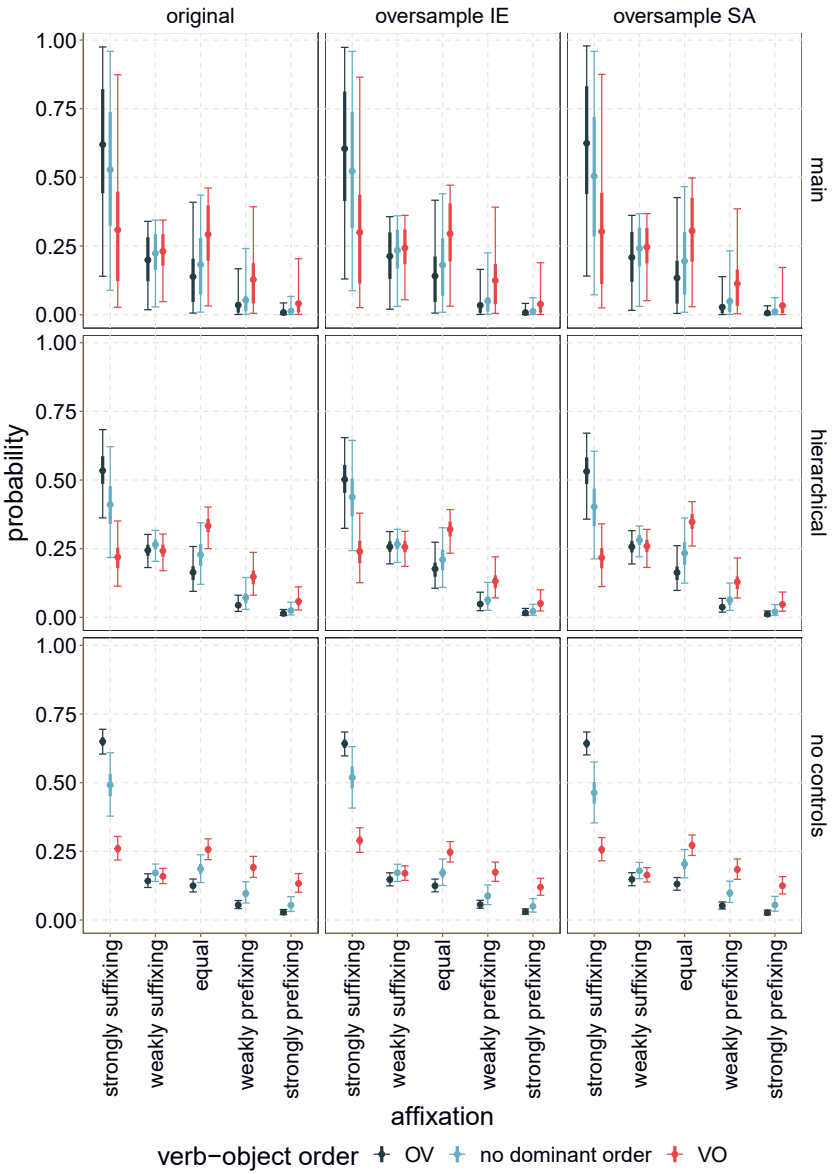

Figure 19: Conditional effects of verb-object order on affix position for oversampled data.

Table 7: ELPD differences for the main, hierarchical, and no-controls models with oversampled data.

		Indo-European		South America	
		ELPD diff	SE diff	ELPD diff	SE diff
1	phylo + areal GP + verb-object	0.0	0.0	0.0	0.0
2	(1|family) + (1|macroarea) + verb-object	−120.4	12.8	−75.6	11.6
3	verb-object	−316.0	16.3	−274.6	15.6

from our findings is that typological samples used to examine crosslinguistic trends should really be as large as possible, potentially larger than the common 100 or 200 languages sample sizes. Of course, more testing with different samples and linguistic features would be needed to yield more conclusive results about this question. Simulating a higher degree of oversampling, we did break our main model by adding a single datapoint an additional 1,000 times (not shown here).

To conclude this section, our claim is not that any set of languages can make a valid sample, but rather that as long as one includes many languages from different genealogical groups and areas, one does not need to worry too much about careful, systematic sampling procedures. Most importantly, our results do not point to any reason for excluding data.

7 Discussion

7.1 The importance of areal and contact bias control

In this paper, we proposed and explored two techniques to control for genealogical and areal biases. Sampling in typology traditionally emphasizes genealogical bias, both when controlling for biases in the sampling process itself and through statistical modeling. Our results, on the other hand, point towards areal effects being an equally important factor to control for as genealogical relations.

Previous studies on borrowing of affixation and word order patterns and on their diachronic stability have shown different results. Usually, word order has been argued to be subject to areal diffusion and thus to be a rather unstable property (Croft et al. 2011: 446). On the other hand, Parkvall (2008), examining various grammatical properties for stability within different language families, finds word order related properties to be amongst the most stable ones (see also Dediu and Cysouw 2013 for further discussion on the stability of linguistic

features). Affixation has long been assumed to be less prone to borrowing. While this assumption has been proven incorrect by many studies over the last 20 years, Seifart (2015) also shows that affix borrowing is not less likely to happen between languages that are typologically very different than between languages that have similar structures. Thus, since our study only examined the effect of word order on affix position, together with the two types of bias control, our results may not extend to all other properties of grammar, given that word order and affix borrowing could be properties that are very susceptible to contact, while other properties may not be.

We would argue that this is rather unlikely. First, as was mentioned, we still lack sufficient conclusive evidence on the relative stability of different grammatical properties.

Second, while areal effects seem to play a background role in sampling (when applied, not in studies that explicitly deal with sampling methods), others have mentioned its importance, also in relation to genealogical bias (cf. also the discussion in Section 4.3.1). Modeling the phonological diversity of languages (unrelated to word order and affix position), Jaeger et al. (2011) include a measure to control for contact bias. Referring to this measure, the authors conclude: "Interestingly, the best weighted areal normalized phonological diversity [=the measure of contact control] (s = 685) seems to capture all relevant continent-level, subfamily-, and genus-level information as well as some country-level information."

Third, it has long been noted in the quantitative typology and sampling literature that contact and areal control are very important to consider (Bickel 2011).

One could argue that we pick up relatively little from the phylogenetic signal (in comparison to the areal signal) because our phylogenetic information is imperfect/incomplete. This is true to a certain extent, and it is possible that a much better phylogenetic structure could improve our model or could turn out to be a more important predictor of affix position. The issues about genealogical control in sampling raised by Dahl (2008: 210) also apply to our method; it requires a complete tree classification of the languages in the sample, and these classifications are not without controversy, especially in the Americas (Dahl 2008: 210). Thus, the accuracy of different branches in the tree is most likely of uneven quality.

However, the phylogenetic structure from Glottolog that we used in this study is arguably the most informed phylogeny currently available. It is possible that a better phylogeny would improve our results and displace areal effects, but areal effects were nevertheless strong in our case (and also in e.g. Jaeger et al. 2011). Thus, even though we cannot be certain without further testing to what extent our results will generalize to other types of linguistic structures, we have solid grounds

to assume that controlling for areal bias is very important for typological studies that examine the crosslinguistic distribution of linguistic features.

The control for areal bias proposed in Section 4.3.2 was inspired by previous theoretical and practical proposals for areal and contact control in Cysouw et al. (2012b), Dryer (2018), and Jaeger et al. (2011). We would argue that our approach combines the relevant insights of these previous studies. Jaeger et al. (2011) is, as far as we are aware, the only quantitative proposal to control for contact bias between neighboring languages via modeling, using the geographical locations of where the languages are spoken. While this measure allows the incorporation of the basic insight that languages that are spoken in geographic proximity influence each other more than geographically more distant languages, the effect of distance in their model is uniform across all areas of the world.

However, as mentioned in Section 4.3.1, work on linguistic diversity has shown that languages are not spread evenly across the globe, meaning that the effect of distance between languages varies across geographical areas. This is also what Cysouw et al. (2012b) and Dryer (2018) have pointed out and what Dryer applied to sampling, proposing a measure of languages spoken in between a pair of languages (cf. Section 4.3.1).

Our approach to capturing areal effects based on geographic information is a refined version of the measure proposed in Jaeger et al. (2011), taking into account Dryer's argument. Like Jaeger et al. (2011), we use distances between languages as the basis for our method and like Dryer's method, ours takes into account the varying effect of distances depending on the geographic area.

7.2 "Static" and "dynamic" approaches to crosslinguistic tendencies

Our approach falls within the traditional way of testing for crosslinguistic tendencies: building a sample and examining the distribution of the phenomenon in question in the languages of the sample. Regardless of the exact sampling criteria and whether or not testing for the distribution is done with statistical modeling, this approach generally estimates the distribution of the phenomenon from the attested distribution in the language sample. As a consequence, typological universals or crosslinguistic tendencies are expressed as a static, distributional statement as well. For instance, one can estimate the probability of a given phenomenon in the world's languages, or the probability of two properties co-occurring, etc. Regarding the test case of the present study, this means that we estimated the association of different word order patterns and affix positions. Once

genealogical and areal biases are factored in, we found no strong association between any word order and affixation pattern.

Greenberg (1978) introduced the idea of what he called diachronic approaches to typology and the idea of transitional probabilities. These ideas were first implemented by Maslova (2000), who proposed a concrete mathematical method for calculating transitional probabilities. This so-called "dynamic" approach has been implemented and discussed in various ways in a number of studies over the last 20 years (Bickel 2013, 2017; Cysouw 2011; Dediu 2011; Dunn et al. 2011; Maslova and Nikitina 2007). In contrast to the traditional "static" approach, the "dynamic" approach is based on the diachronic changes between values of the feature(s) in question across related languages in order to estimate the probability of one feature value changing into another feature value. In this approach, typological universals are no longer expressed in terms of the proportions of languages that have a certain feature value, but in terms of the probability of a change from one feature value to another.

While dynamic typology presents an interesting alternative to the more traditional "static" approach, the way in which transition probabilities have been used so far to uncover crosslinguistic trends and patterns is potentially problematic. To the best of our knowledge, many studies following this approach focus on precise information about genealogical relations at the expense of contact and areal control (although see Bickel 2013 for a method that can take, to a limited degree, areal patterns into consideration). It could be the case that for a certain phenomenon, contact and areal factors does not play much of a role in the relevant transitions over time. However, our results, as well as results from previous studies, have shown that this is very unlikely. Thus, comparing transition probabilities in a rather small number of phyla as in Dunn et al. (2011) could lead to a bias in the estimated transition probabilities because of contact or areal factors that interact with the genealogical relations between languages. If these are not controlled for, we may end up with transition probabilities biased by contact and diffusion that can distort our interpretation of transition probabilities estimated on the basis of genealogical trees. Besides other issues raised by Cysouw (2011), where he explains why Dunn et al. (2011) should be more careful in claiming that they have presented evidence against a general crosslinguistic trend of word order, the lack of controlling for contact and areal effects could also lead to different distributions in different families. For instance, finding different transition probabilities across different families without controlling for contact and areal effects cannot be taken as evidence for the lack of a more general trend, since different families

could additionally be impacted by different areal factors.[42] The opposite is also true: not controlling for areal effects could also lead to falsely finding very similar transitional probabilities across different families.

Another important point is that traditional "static" approaches can handle isolate languages much better than "dynamic" approaches using transition probabilities. Isolates such as Basque are languages that cannot be related with certainty to a known language family. Of course, this reflects our knowledge about genealogical relations and existing language families, and isolates could have belonged to a larger language family in the past that we simply do not know of. Isolates are crosslinguistically fairly common. Campbell (2010: 28) notes that about one third of all language families have only a single member, i.e. an isolate. While the details may differ according to the sampling method used, nothing prevents us from including isolates in a traditional sample used to analyze synchronic distributions.[43] Using transition probabilities, on the other hand, relies on our knowledge about the genealogical relations between languages. This is because transition probabilities from one feature value to another are estimated by the comparison of (closely) related languages. This issue has not received sufficient attention in "dynamic" methods to examine language universals, one notable exception being the Family Bias Method (Bickel 2011, 2013). Bickel proposes to estimate transition probabilities in small language families and isolates by extrapolation of the bias in large language families. While it is certainly possible to extend models to estimate transition probability used in dynamic typology so that isolates can be included, it is not straightforward and requires sufficient data from large families to build on.

Although being more similar to the traditional "static" approaches to sampling and language universals, the method presented in this study aims at combining insights from both perspectives. We showed that our method allows for the inclusion of languages regardless their genealogical relations, which is an advantage given the pervasiveness of isolate languages and unclear genealogical relations. Our method results in statements about crosslinguistic distributions in the traditional, "static" way. Yet, as we include controls for genealogical and areal relations between the languages of the sample, our model is not blind to the structure between the languages. Conceptually, our method makes use of closely related languages in a similar way as in the dynamic approaches: whether or not closely related languages have different or similar properties will influence the

42 See Bickel (2017) for the distinction between "functional" (i.e. general) and "event-based" (i.e. areal) triggers of change which influence the distribution of linguistic features.

43 This also applies to sign languages or creoles, whose genealogical classification in terms of traditional language families may not always be trivial.

model estimates and reflect to what extent any associations result from family-specific biases in the sample. Thus, we do not estimate transition probabilities in an explicit way, but implicitly; but the phylogenetic relations between all languages in the sample are included in our model estimating the distribution of features. In fact, as we showed in Section 5, once we include this information, the association between word order and affixation preference becomes very weak. Importantly, our approach does not make any assumptions about how crosslinguistic patterns come about, which makes it perfectly compatible with a "dynamic" interpretation of universals.

In addition, we also control for the effects of areal bias in a (conceptually) similar way in that we take into account that neighboring languages are likely to influence each other. To the best of our knowledge, this has not yet been implemented in any dynamic approaches to universals.

This is not to argue for or against either approach to crosslinguistic distributions and universals. On the contrary, we think that it is important to explore and compare different methods. But we also want to highlight that, no matter which approach is used, additional contact and areal controls besides the phylogenetic control are crucial.

7.3 Bias control: the next steps

We believe that the model presented in this paper is a clear improvement on previous methods of bias control in typology. Nevertheless, there are a number of aspects that still need to be improved on. The first and main issue is that we currently represent the location of languages as single point estimates. This is an insufficient representation in two ways. First, we know that the geographical location of languages would better be captured by polygons representing the area instead of single points with single coordinates for latitude and longitude. Compiling such data is very resource-intensive and unrealistic without the wide-scale collaboration of many scientists. As of now, we are not aware of any freely available crosslinguistic database containing more detailed information on the area that languages are spoken in.[44] A possible alternative would be to transform the point-wise estimates into discrete polygons using Voronoi diagrams (Aurenhammer 1991), as also proposed by Hammarström and Güldemann (2014) and Kälin (2017). However, these approaches are not without issues either, the most salient one being that point-to-polygon conversion is heavily dependent on the

44 Resources such as the World Language Mapping System (https://www.worldgeodatasets.com/language/) offer polygon information but are not freely accessible.

number of languages in the sample. Thus, as far as we can tell, this method is not stable in the sense that for two different samples, the same language L may be assigned to drastically different polygons.

Second, by using single point estimates for the location of languages, we miss language contact involving a lingua franca spoken sufficiently far away from the point estimate. Of course it would be important to test for the possibility that, for instance, Spanish could have an influence on Nahuatl, or that French could influence Wolof.

In addition to these evident geographical simplifications that we currently have to make, we would ideally like to be able to represent language contact in a much more realistic way in terms of socio-linguistic and geological variables that play an important role in shaping language contact situations. As was mentioned in Section 4.3.1, there are various socio-linguistic factors that determine if and how language contact takes place. Often, contact that leads to borrowing of grammatical structures is not necessarily bidirectional, but instead unidirectional from the more dominant to the less dominant language. Different sociological factors such as the political status of the languages, the language attitude of the speakers, and the communal level of multilingualism also determine how likely it is to borrow patterns from another language. Thus, geographic proximity may often (but not always) lead to language contact, which could then in turn lead to contact-induced language change.

In its current state, our model is fairly naive with respect to the geographic and geological realities. It does not include information on the geological properties such as mountain ranges, rivers, etc. However, we know that such geological properties have influenced the movement and migration patterns of people and thus the probabilities of language contact as well (van Gijn et al. 2017). Including this type of information systematically and for the whole globe is not a trivial endeavor, but it would offer a more realistic perspective on areal effects.

Another possibility which we do not explore in this paper is the idea of building variance models. A variance model does not assume that the variance is equal for all observations, but instead allows certain groups to have different amounts of variance. Thus, one could consider that certain lineages might exhibit less or higher variance, or that the variance of high diversity regions (i.e. regions with many different languages) might be greater than that of low diversity regions.

Another potential line of future research would be to include phylogenetic uncertainty into the models. In this paper, we assume that the phylogenetic trees are fundamentally correct and static. However, this assumption is likely wrong, and using a posterior sample of phylogenetic trees might be more appropriate.

7.4 Inclusive sampling

Another crucial finding of the present paper is that neither the exclusion of languages from a sample nor a highly controlled *a priori* sampling of languages are necessary for analyzing crosslinguistic distributions. With the statistical bias controls proposed in Section 4, all the language data that the researcher has access to can be included in the sample. Of course, in order to capture crosslinguistic trends across all macroareas, one still needs to include languages from as many areas and families as possible. The crucial point, however, is that statistical genealogical and areal control no longer require the researcher to artificially restrict the number of languages that are related or spoken in close proximity to each other.

The possibility to include as many languages as possible that may not be independent from each other has another practical advantage. As mentioned in Section 3.3, sampling methods as well as the resulting language samples can roughly be divided into probability samples and variety samples. In practice, such a distinction would mean that typological studies can either explore the attested values and distributions of a given linguistic feature, or they can examine the distribution in a balanced sample, controlling for genetic, areal, and contact biases to find out about the functional (and also extra-linguistic) factors that influence its distribution. It may be the case that some studies only pursue one of the two objectives, but a more likely scenario is that the same sample will serve as the basis for a first exploration of the phenomenon and a quantitative assessment of the relevant distributions in a second step. Thus, in practice, it seems more useful to be able to build a sample that allows for both tasks.

8 Concluding remarks

In this paper we revisited the question of bias control in typology. To that end, we presented two advanced statistical tools for genealogical and areal control. Using the WALS data for verb-object order and affix position as a test case, we showed how our method can be applied to an unbalanced sample of languages. Contrary to a number of previous studies, our findings pointed to a weak, if not non-existent, effect of verb-object order on affix position, when genealogical relations and particularly areal effects are controlled for (of course, this does not invalidate other linguistic factors that have been found to condition affix position). Model comparison with a hierarchical control for areal and genealogical effects, as well as with a no-controls model, showed that more traditional methods for typological

modeling vastly overestimate the effect of verb-object on affix position and that they fail to properly control for genealogical and areal biases. Since this is in line with the results of other studies examining unrelated phenomena, it is very likely that our findings regarding the role of language contact can be generalized to other linguistic phenomena as well. It is especially surprising given that we modeled language contact in a fairly simple way. As we see it, this has two important consequences for typologists. From a modeling perspective, it should be the goal of quantitative typology to come up with a better representation of language contact in the future. From a more general perspective, our findings show that contact control needs to play a more prominent role in any typological study that makes use of a language sample to examine the distribution of linguistic features.

Author contributions: M.G.N: initial idea and statistical modeling, evaluation of results, writing; L.B.: background and overview of previous typological work, evaluation of results, writing.

References

Aikhenvald, Alexandra Y. 2006. Grammars in contact: A cross-linguistic perspective. In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *Grammars in contact: A cross-linguistic typology*, 1–66. Oxford: Oxford University Press.

Aikhenvald, Alexandra Y. & R. M. W. Dixon. 2006a. *Areal diffusion and genetic inheritance: Problems in comparative linguistics*. Oxford: Oxford University Press.

Aikhenvald, Alexandra Y. & R. M. W. Dixon. 2006b. *Grammars in contact: A cross-linguistic typology*. Oxford: Oxford University Press.

Aurenhammer, Franz. 1991. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)* 23(3). 345–405.

Baayen, Harald & Maja Linke. to appear. An introduction to the generalized additive model. In Stefan Gries & Magali Paquot (eds.), *A practical handbook of corpus linguistics*. Berlin: Springer.

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

Becker, Laura. 2021. *Articles in the world's languages* (Linguistische Arbeiten 577). Berlin: De Gruyter.

Bell, Alan. 1978. Language samples. In Joseph H. Greenberg & Charles Albert Ferguson (eds.), *Universals of human language. Volume 1: Method and theory*, 123–156. Stanford, CA: Stanford University Press.

Bentz, Christian, Annemarie Verkerk, Douwe Kiela, Felix Hill & Paula Buttery. 2015. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLOS ONE* 10(6). e0128254.

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

Bickel, Balthasar. 2008. A refined sampling procedure for genealogical control. *Language Typology and Universals* 61(3). 221–233.
Bickel, Balthasar. 2011. Statistical modeling of language universals. *Linguistic Typology* 15(2). 401–413.
Bickel, Balthasar. 2013. Distributional biases in language families. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency* (Typological Studies in Language 104), 415–444. Amsterdam: John Benjamins.
Bickel, Balthasar. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Narrog Heiko (eds.), *Oxford handbook of linguistic analysis*, 2nd edn. (Oxford Handbooks in Linguistics). Oxford: Oxford University Press.
Bickel, Balthasar. 2017. Areas and universals. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics* (Cambridge Handbooks in Language and Linguistics), 40–55. Cambridge: Cambridge University Press.
Bickel, Balthasar & Johanna Nichols. 2006. Oceania, the Pacific Rim, and the theory of linguistic areas. *Annual Meeting of the Berkeley Linguistics Society* 32(2). 3–15.
Bickel, Balthasar & Johanna Nichols. 2013. The Autotyp genealogy and geography database 2013 release. Available at: http://www.autotyp.uzh.ch/available.html.
Bisang, Walter, Laura Becker, Andrej Malchukov & Marvin Martiny. 2020. Grammaticalization scenarios: Constraining typological variation. Submitted for publication.
Blasi, Damián, Steven Moran, Scott R. Moisik, Paul Widmer, Dan Dediu & Balthasar Bickel. 2019. Human sound systems are shaped by post-neolithic changes in bite configuration. *Science* 363(6432). eaav3218.
Bouckaert, Remco R., Claire Bowern & Quentin D. Atkinson. 2018. The origin and expansion of Pama-Nyungan languages across Australia. *Nature Ecology & Evolution* 2(4). 741–749.
Bowern, Claire & Quentin D. Atkinson. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88(2). 817–845.
Bürkner, Paul-Christian. 2017. Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–28.
Bürkner, Paul-Christian. 2018. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* 10(1). 395–411.
Bybee, Joan L., William Pagliuca & Revere Dale Perkins. 1990. On the asymmetries in the affixation of grammatical material. In William Croft, Suzanne Kemmer & Denning Keith (eds.), *Studies in typology and diachrony. Papers presented to Joseph H. Greenberg on his 75th birthday*, 1–42. Amsterdam: Benjamins.
Bybee, Joan L., Revere Dale Perkins & William Pagliuca. 1994. *The evolution of grammar*. Chicago: The University of Chicago Press.
Campbell, Lyle. 2010. Language isolates and their history, or, what's weird, anyway? *Annual Meeting of the Berkeley Linguistics Society* 36(1). 16–31.
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, Articles* 76(1). 1–32.
Comrie, Bernard, Matthew S. Dryer, David Gil & Martin Haspelmath. 2013. Introduction. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Croft, William, Tanmoy Bhattacharya, Dave Kleinschmidt, D. Eric Smith & T. Florian Jaeger. 2011. Greenbergian universals, diachrony, and statistical analyses. *Linguistic Typology* 15(2). 433–453.

Cutler, Anne, John A. Hawkins & Gary Gilligan. 1985. The suffixing preference: A processing explanation. *Linguistics* 23. 723–758.

Cysouw, Michael. 2005. Quantitative methods in typology. In *Quantitative Linguistik/Quantitative Linguistics* (HSK 27), 554–578. Berlin: De Gruyter.

Cysouw, Michael. 2010. Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14(2–3). 253–286.

Cysouw, Michael. 2011. Understanding transition probabilities. *Linguistic Typology* 15(2). 415–431.

Cysouw, Michael, Dan Dediu & Steve Moran. 2012a. Comment on "Phonemic diversity supports a serial founder effect model of language expansion from Africa". *Science* 335(6069). 657.

Cysouw, Michael, Dan Dediu & Steven Moran. 2012b. Supporting online material for: Comment on "Phonemic diversity supports a serial founder effect model of language expansion from Africa". https://science.sciencemag.org/content/sci/suppl/2012/02/09/335.6069.657-b.DC1/Cysouw.SOM.pdf (accessed 27 August 2020).

Dahl, Östen. 2001. Principles of areal typology. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals*, vol. 2, 1456–1470. Berlin: De Gruyter.

Dahl, Östen. 2008. An exercise in a posteriori language sampling. *Language Typology and Universals* 61(3). 208–220.

Dediu, Dan. 2011. A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proceedings of the Royal Society B: Biological Sciences* 278(1704). 474–479.

Dediu, Dan & Michael Cysouw. 2013. Some structural aspects of language are more stable than others: A comparison of seven methods. *PLOS ONE* 8(1). e5500.

Dediu, Dan & Stephen C. Levinson. 2012. Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLOS ONE* 7(9). e45198.

de Villemereuil, Pierre & Shinichi Nakagawa. 2014. *Modern phylogenetic comparative methods and their application in evolutionary biology*. Berlin: Springer.

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

Donohue, Mark & Bronwen Whiting. 2011. Quantifying areality: A study of prenasalisation in Southeast Asia and New Guinea. *Linguistic Typology* 15(1). 101–121.

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

Dryer, Matthew S. 1991. SVO languages and the OV: VO typology. *Journal of Linguistics* 27(2). 443–482.

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68. 81–138.

Dryer, Matthew S. 2011. The evidence for word order correlations. *Linguistic Typology* 15(2). 335–380.

Dryer, Matthew S. 2013a. Order of object and verb. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: https://wals.info/chapter/83.

Dryer, Matthew S. 2013b. Prefixing vs. suffixing in inflectional morphology. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: https://wals.info/chapter/26.

Dryer, Matthew S. 2018. On the order of demonstrative, numeral, adjective and noun. *Language* 94(4). 798–833.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: https://wals.info/.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473. 79–82.

Enfield, Nick. 2005. Areal linguistics and mainland Southeast Asia. *Annual Review of Anthropology* 34(1). 181–206.

Enrique-Arias, Andrés. 2002. Accounting for the position of verbal agreement morphology with psycholinguistic and diachronic explanatory factors. *Studies in Language* 26(1). 1–31.

Faraway, Julian J. 2006. *Extending the linear model with R*. London: Taylor & Francis.

Foster, Joseph F. & Charles A. Hofling. 1987. Word order, case, and agreement. *Linguistics* 25(3). 475–500.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt & Andrew Gelman. 2019. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182(2). 389–402.

Garland, Theodore Jr. & Anthony R. Ives. 2000. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *The American Naturalist* 155(3). 346–364.

Gaudette Lisa, Japkowicz Nathalie. 2009. Evaluation methods for ordinal classification. In Gao Yong, Japkowicz Nathalie (eds.), *Advances in artificial intelligence. Canadian AI 2009. Lecture notes in computer science*, vol. 5549, Springer, Berlin, Heidelberg.

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gelman, Andrew & Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Unpublished manuscript.

Gelman, Andrew & Eric Loken. 2014. The statistical crisis in science: Data-dependent analysis-a "garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American Scientist* 102(6). 460–466.

Gray, Russell D., Alexei Drummond & Simon Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323. 479–483.

Greenberg, Joseph H. 1978. Diachrony, synchrony and language universals. In Joseph H. Greenberg, Charles Albert Ferguson & Edith A. Moravcsik (eds.), *Universals of human language*, vol. 1: Method & Theory, 61–92. Standford: Standford 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. 2020. *Glottolog 4.3*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Hammarström, Harald & Tom Güldemann. 2014. Quantifying geographical determinants of large-scale distributions of linguistic features. *Language Dynamics and Change*. 4(1). 87–115.

Hawkins, John A. & Anne Cutler. 1988. Psycholinguistic factors in morphological asymmetry. In John A. Hawkins (ed.), *Explaining language universals*, 280–317. Oxford: Blackwell.

Hawkins, John A. & Gary Gilligan. 1988. Prefixing and suffixing universals in relation to basic word order. *Lingua* 74. 219–259.
Hetterle, Katja. 2015. *Adverbial clauses in cross-linguistic perspective*. Berlin: De Gruyter.
Hickey, Raymond. 2010. *The handbook of language contact*. Malden, MA: Wiley-Blackwell.
Hickey, Raymond (ed.). 2017. *The Cambridge handbook of areal linguistics* (Cambridge Handbooks in Language and Linguistics). Cambridge: Cambridge University Press.
Himmelmann, Nikolaus P. 2000. Towards a typology of typologies. *Language Typology and Universals* 53(1). 5–12.
Himmelmann, Nikolaus P. 2014. Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language* 90(4). 927–960.
Holman, Eric W., Christian Schulze, Dietrich Stauffer & Søren Wichmann. 2007. On the relation between structural diversity and geographical distance among languages: Observations and computer simulations. *Linguistic Typology* 11(2). 393–421.
Housworth, Elizabeth A., Emilia P. Martins & Michael Lynch. 2004. The phylogenetic mixed model. *The American Naturalist* 163(1). 84–96.
Hua, Xia, Simon J. Greenhill, Marcel Cardillo, Hilde Schneemann & Lindell Bromham. 2019. The ecological drivers of variation in global language diversity. *Nature Communications* 10(1). 1–10.
Jaeger, T. Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15(2). 281–319.
Jäger, Gerhard. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* 3. 245–291.
Jäger, Gerhard. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data* 5(1). 1–16.
Jäger, Gerhard. 2019. Computational historical linguistics. *Theoretical Linguistics* 45(3–4). 151–182.
Jäger, Gerhard & Johanes Wahle. Forthcoming. *Phylogenetic typology*. https://www.researchgate.net/publication/350160257_Phylogenetic_typology (accessed 1 June 2021).
Kälin, Fabiola. 2017. *Global analysis of the influence of geographical factors on contact-induced language change*. Zürich: Geographisches Institut der Universität Zürich.
Kimeldorf, George S. & Grace Wahba. 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* 41(2). 495–502.
Koch, Harold. 2014. Historical relations among the Australian languages: Genetic classification and contact-based diffusion. In Harold Koch & Rachel Nordlinger (eds.), *The languages and linguistics of Australia: A comprehensive guide* (The World of Linguistics), 23–89. Berlin: De Gruyter.
Levinson, Stephen C., Simon J. Greenhill, Russell D. Gray & Michael Dunn. 2011. Universal typological dependencies should be detectable in the history of language families. *Linguistic Typology* 15(2). 509–534.
Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533–572.
van Lier, Eva. 2016. Lexical flexibility in Oceanic languages. *Linguistic Typology* 20(2). 197–232.
List, Johann-Mattis. 2019. Automated methods for the investigation of language contact, with a focus on lexical borrowing. *Language and Linguistics Compass* 13(10). e12355.

List, Johann-Mattis, Nelson-Sathi Shijulal, William Martin & Hans Geisler. 2014. Using phylogenetic networks to model Chinese dialect history. *Language Dynamics and Change* 4. 222–252.

Louagie, Dana & Jean-Christophe Verstraete. 2016. Noun phrase constituency in Australian languages: A typological study. *Linguistic Typology* 20(1). 25–80.

Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLOS ONE* 5(1). e8559.

Martowicz, Anna. 2011. *The origin and functioning of circumstantial clause linkers: A cross-linguistic study*. PhD dissertation, University of Edinburgh.

Maslova, Elena. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4(3). 307–333.

Maslova, Elena & Tatiana Nikitina. 2007. Stochastic universals and dynamics of cross-linguistic distributions: The case of alignment type. Unpublished manuscript. Available at: http://anothersumma.net/Publications/Ergativity.pdf.

Matras, Yaron & Jeanette Sakel (eds.). 2008. *Grammatical borrowing in cross-linguistic perspective*. Berlin: De Gruyter.

Maurits, Luke & Thomas L. Griffiths. 2014. Tracing the roots of syntax with Bayesian phylogenetics. *Proceedings of the National Academy of Sciences* 111(37). 13576–13581.

Miestamo, Matti. 2003. *Clausal negation: A typological study*. Helsinki: Helsingin Yliopisto.

Miestamo, Matti. 2005. *Standard negation: The negation of declarative verbal main clauses in a typological perspective* (Empirical Approaches to Language Typology 31). Berlin: De Gruyter.

Miestamo, Matti, Dik Bakker & Antti Arppe. 2016. Sampling for variety. *Linguistic Typology* 20(2). 233–296.

Murawaki, Yugo. 2015. Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 324–334. Denver: Association for Computational Linguistics.

Murawaki, Yugo. 2018. Analyzing correlated evolution of multiple features using latent representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4371–4382. Brussels: Association for Computational Linguistics.

Murawaki, Yugo & Kenji Yamauchi. 2018. A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. *Journal of Language Evolution* 3(1). 13–25.

Murdock, George Peter. 1967. *Ethnographic atlas*. Pittsburgh, PA: University of Pittsburgh Press.

Nettle, Daniel. 1999. Is the rate of linguistic change constant? *Lingua* 108(2). 119–136.

Nichols, Johanna. 1986. Head-marking and dependent-marking grammar. *Language* 62(1). 56–119.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Nikolaev, Dmitry & Eitan Grossman. 2018. Areal sound change and the distributional typology of affricate richness in Eurasia. *Studies in Language* 42(3). 562–599.

Pagel, Mark. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255(1342). 37–45.

Parkvall, Mikael. 2008. Which parts of language are the most stable? *Language Typology and Universals* 61(3). 234–250.

Perkins, Revere Dale. 1980. *The evolution of culture and grammar*. New York: State University of New York at Buffalo.
Perkins, Revere Dale. 1989. Statistical techniques for determining language sample size. *Studies in Language* 13(2). 293–315.
R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna, Austria: Manual.
Rasmussen, Carl Edward. 2003. Gaussian processes in machine learning. In Olivier Bousquet, Ulrike von Luxburg & Gunnar Rätsch (eds.), *Advanced lectures on machine learning. ML 2003. Lecture notes in computer science*, vol. 3176, 63–71. Berlin: Springer.
Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2(3). 263–314.
Rijkhoff, Jan, Dik Bakker, Kees Hengeveld & Peter Kahrel. 1993. A method of language sampling. *Studies in Language* 17(1). 169–203.
Schmidtke-Bode, Karsten. 2009. *A typology of purpose clauses* (Typological Studies in Language 88). Amsterdam: John Benjamins.
Seifart, Frank. 2015. Does structural-typological similarity affect borrowability?: A quantitative study on affix borrowing. *Language Dynamics and Change* 5(1). 92–113.
Siemund, Peter & Noemi Kintana (eds.). 2008. *Language contact and contact languages*. Amsterdam: John Benjamins.
Siewierska, Anna & Dik Bakker. 1996. The distribution of subject and object agreement and word order type. *Studies in Language* 20. 115–161.
Sinnemäki, Kaius. 2010. Word order in zero-marking languages. *Studies in Language* 34(4). 869–912.
Sinnemäki, Kaius. 2014. A typological perspective on differential object marking. *Linguistics* 52(2). 281–314.
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. https://doi.org/10.3389/fpsyg.2018.01141.
Song, Jae Jung. 2012. *Word order*. Cambridge: Cambridge University Press.
Song, Jae Jung. 2018. *Linguistic typology*. Oxford: Oxford University Press.
Stassen, Leon. 1985. *Comparison and universal grammar*. Oxford: Blackwell.
Steele, Susan. 1978. Word order variation: A typological survey. In Joseph Harold Greenberg, Charles Albert Ferguson & Edith A. Moravcsik (eds.), *Universals of human language IV: Syntax*, 585–623. Stanford, CA: Stanford University Press.
Thomason, Sarah Grey. 2001. *Language contact: An introduction*. Washington, D.C.: Georgetown University Press.
Urban, Matthias, Hugo Reyes-Centeno, Kate Bellamy & Matthias Pache. 2019. The areal typology of western Middle and South America: Towards a comprehensive view. *Linguistics* 57(6). 1403–1463.
van Gijn, Rik, Harald Hammarström & Simon van de Kerke. 2017. Linguistic areas, linguistic convergence and river systems in South America. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics* (Cambridge Handbooks in Language and Linguistics), 964–996. Cambridge: Cambridge 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.
Vehtari, Aki, Daniel Simpson, Andrew Gelman, Yuling Yao & Jonah Gabry. 2021. Pareto smoothed importance sampling. Unpublished manuscript. Available at: https://arxiv.org/abs/1507.02646.
Verkerk, Annemarie. 2019. Detecting non-tree-like signal using multiple tree topologies. *Journal of Historical Linguistics* 9(1). 9–69.
Voegelin, Charles Frederick & Florence Marie Voegelin. 1977. *Classification and index of the world's languages*. New York: Elsevier.
Wichmann, Søren & Harald Hammarström. 2020. Methods for calculating walking distances. *Physica A: Statistical Mechanics and its Applications* 540. 122890.
Wieling, Martijn, John Nerbonne & R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLOS ONE* 6(9). e23613.
Williams, Christopher K. I. & Carl Edward Rasmussen. 2006. *Gaussian processes for machine learning*, vol. 2. Cambridge, MA: MIT Press.
Ye, Jingting. 2020. *Property words and adjective subclasses in the world's languages*. PhD dissertation, Leipzig University.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.5281/zenodo.5576242).