A typological approach to language change in contact situations

Kaius Sinnemäki, Francesca Di Garbo, Ricardo Napoleão de Souza, T. Mark Ellison

A typological approach to language change in contact situations

Kaius Sinnemäki[1], Francesca Di Garbo[1,2],
Ricardo Napoleão de Souza[1,3] and T. Mark Ellison[4]
[1] University of Helsinki | [2] Aix-Marseille University, CNRS, LPL |
[3] University of Edinburgh | [4] University of Cologne

Language contact phenomena have increasingly been researched from different historical linguistic, sociolinguistic and areal-typological perspectives. However, since most of this research is based on case studies, an assessment of contact phenomena from a worldwide comparative perspective has been missing in the literature. In this article, we draw inspiration from historical linguistics and language typology to present a new typological approach for evaluating evidence that given linguistic domains have been affected by language contact. This method has three parts: (1) a new approach to sampling, (2) the analysis of typological data, and (3) making probabilistic inferences about language contact. We argue that this is a parsimonious method for evaluating contact effects that can serve as a starting point for the further development of typological approaches to language contact.

Keywords: language contact, language typology, language variation, sampling, convergence, stability, Bayesian approach

1. Introduction

There has long been a debate in historical linguistics regarding language-internal versus language-external motivations for linguistic change. In that debate, language-external motivations, mostly related to language contact, have arguably been viewed less favourably. Some of the criticisms that have been directed at contact explanations include a lack of awareness of confounding factors, especially universal preferences (e.g. Ranacher et al. 2021), issues in generalising findings from individual case studies (e.g. Backus 2014), and a disregard for the sociolinguistic contexts in which contact takes place (e.g. Yakpo 2020).

https://doi.org/10.1075/dia.23029.sin | Published online: 25 June 2024
Diachronica ISSN 0176-4225 | E-ISSN 1569-9714

On the other hand, the vast literature on contact and areal linguistics demonstrates that language-external pressures do exert an impact on language structures, one which language-internal phenomena alone may fall short of explicating. Earlier research on language contact has emphasised, for instance, how contact affects reconstruction within families (e.g. Bowern 2013), how it gives rise to linguistic areas (e.g. Ranacher et al. 2021), or how to classify outcomes of contact in terms of the sociolinguistic context (e.g. Trudgill 2011; Croft 2021). Still, much of what is known about contact effects derives from case studies. While case studies illuminate possible types of change, they rarely allow for an assessment of the likelihood of contact effects across languages. To better understand the nature of language change in contact situations, research needs to compare linguistic structure across diverse situations.

A systematic typological approach to evaluating language contact may offer a solution to these shortcomings (e.g. Koptjevskaja-Tamm 2010: 569–570). Nonetheless, there are currently no methods in typology geared towards uncovering recurrent cross-linguistic pathways of change in contact situations. In this article, we describe a typological approach to language contact that lays the foundation for addressing these issues. The method can be further developed to accommodate the different rates at which linguistic features change across languages; that is, universal tendencies in language change (see Supplement S4). Finally, the method can be expanded to incorporate sociolinguistic and geographic data. This can be done by applying the sampling scheme illustrated in §2 to sociolinguistic and demographic data collection (cf. Thomason & Kaufman 1988; Di Garbo et al. 2021; Napoleão de Souza et al. 2022).

In our approach, we build on tried and tested methods from language typology, further developing them to suit research on language contact. Typology has largely focused on investigating general questions about cross-language distributions, such as finding statistical language universals. However, a typological approach to language contact calls for a reassessment of these methods to one that is geared towards researching contact from the outset. This overall vision has repercussions for sampling units, for the selection of linguistic features to analyse, and for the inferences we make about contact effects. Our design addresses all of these aspects through three main components: (1) a novel approach to sampling for contact, (2) the typological analysis of linguistic variables, and (3) a probabilistic evaluation of evidence for contact in the data.

Our sampling units are pairs of languages that have been in contact with one another, along with a third language that serves as a control for inheritance factors (§2). Furthermore, we argue that language-internal variation holds the key to a better understanding of how languages change in contact situations (§3). In order to capture language-internal variation, we adopt principles of multivariate typol-

ogy (e.g. Bickel 2010), which are designed to research such variation in typology. Finally, we propose to evaluate evidence of contact using several linguistic categories, such as nominal number, adnominal possession, or syllable structure. That is, we analyse a number of language-internal features in each of those categories and then draw conclusions by aggregating evidence later in the data analyses (cf. Witzlack-Makarevich et al. 2022). We then illustrate how evidence for language contact can be assessed within a Bayesian framework (§4). In §5, we conclude the article with a brief discussion of the implications of this approach.

2. Sampling for contact

Sampling is extensively debated in typology. Large-scale comparative investigations of language structures depend on which languages one selects, and how that selection represents the population of languages of the world (Miestamo et al. 2016). The increasing interest in diachronic approaches to typological research has raised discussions about sampling methods that would enable studying genealogical diversity through phylogenetic comparative methods (Macklin-Cordes & Round 2022) or other quantitative methods (e.g. Maslova 2003). Research that tackles areality and genealogy in typological distributions directly addresses the issue of representativeness and independence (e.g. Cathcart et al. 2018 on Indo-European; Guzmán Naranjo & Becker 2022 on worldwide sampling). However, discussions of sampling methods for language contact research are much rarer. Torres Cacoullos and Travis (2018: Chapter 6) briefly discuss methodological issues related to comparative language contact research within a sociolinguistic variationist approach. Polinsky's (2014) research on heritage languages implements an approach to researching language contact that is conceptually similar to what we propose here for language typology.

Against this background, our sampling scheme introduces a method for global comparisons of language contact scenarios (Di Garbo & Napoleão de Souza 2023). The underlying assumption is that two levels of analysis are needed to infer contact-induced change from large typological datasets: (1) comparing sets of languages in contact and (2) assessing the probability of change against an external measure of control. This twofold approach reflects the widely accepted claim that control data are needed to anchor contact explanations to historical processes (Thomason 2001).

Ideally, the analysis should be based on historical data from different points in time (e.g. Thomason 2001; Torres Cacoullos & Travis 2018). However, historical data have survived mostly from the languages of Europe. Relying solely on historical corpora would thus introduce a significant (Indo-)European bias in typo-

logical datasets for language contact research. The alternative of reconstructing ancestral states via computational phylogenetic methods, while widely adopted in computational historical linguistics, has been used mostly for inferring subgroupings, and rarely for reconstructions of structural features (e.g. Jäger & List 2018). Yet, it is also unclear how many languages per family are required to attain plausible reconstructions using phylogenetic methods, although a minimum of 15–20 languages has been used earlier (e.g. Dunn et al. 2011; Bickel et al. 2015). More importantly, methods for ancestral state reconstruction do not assume any external influence, which makes them less suitable for the purpose of researching contact (but see List 2019; Neureiter et al. 2022; Hübler 2022 for some advances in this regard).

To circumvent these challenges, Di Garbo and Napoleão de Souza (2023) present a proposal that enables genealogical control in global samples. Languages are sampled in sets of three: (i) the Focus language, which is evaluated for contact effects, (ii) the Neighbour language, which is genealogically unrelated to the Focus language but identified as the potential source of linguistic influence on it, and (iii) the Benchmark language, a relative of the Focus language that has not been in contact with either language and which serves as a control to disentangle contact effects from inheritance in the Focus language. One example from Southeast Asia is illustrated in Figure 1.

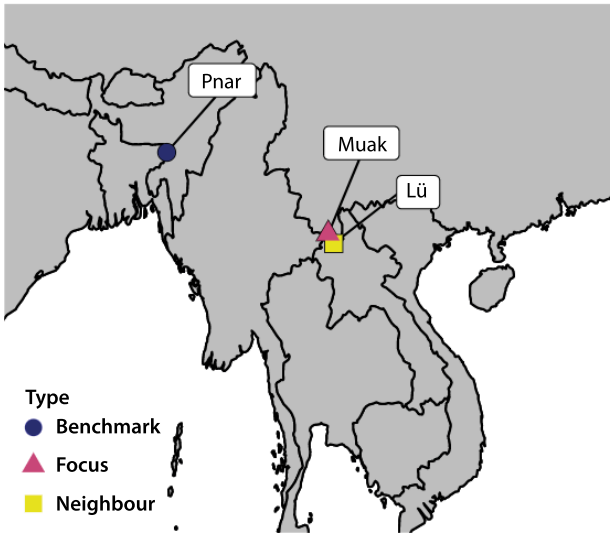


Figure 1. Map illustrating Muak as the Focus language (pink triangle on the map), Pnar as the Benchmark language (dark blue dot) and Lü as the Neighbour language (yellow square)

Our dataset currently consists of 49 sets of three languages, for a total of 147 languages. These have been selected using the 24 AUTOTYP areas as a grid for zooming in on wider continental areas (Bickel et al. 2022). Existing linguistic, historical and anthropological literature was used to further establish candidate contact scenarios for each of these areas. Crucially, pre-existing research on contact-induced change in specific domains of grammar played no role in language selection. Rather, the paramount factors when establishing candidate triplets were: (a) the lack of genealogical links between candidate Focus and Neighbour languages, (b) the lack of contact between those languages and the candidate Benchmark, and (c) the existence of adequate descriptive materials about each language. For a detailed description, see Di Garbo and Napoleão de Souza (2023). Figure 2 shows the geographical distribution of all 147 languages included in the sample.

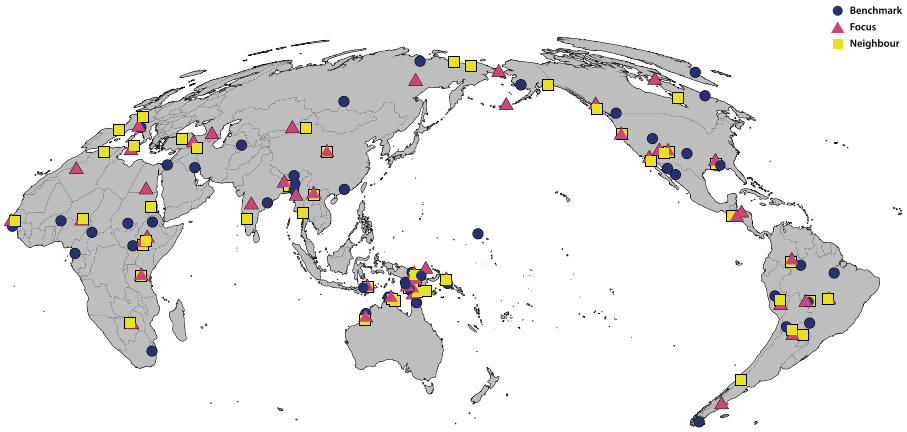


Figure 2. The language sample as in Di Garbo and Napoleão de Souza (2023). The pink triangles represent the Focus languages, the dark blue dots the Benchmark languages, and the yellow squares the Neighbour languages

In sum, this sampling method implements the idea of control data in comparative contact research in a parsimonious way. It also introduces a shift from how sampling units are conceived by existing sampling methods in language typology, which typically use one language as the sampling unit. Given our focus on comparing contact dynamics and their potential effects on language structures, our unit of analysis necessarily consists of minimally three languages – although the method is expandable to include more Benchmark languages per set. This sampling method results in synchronic snapshots of similarities between languages in contact (the Focus and Neighbour), with an added degree of historical control (the Benchmark). The method thus lies in between traditional approaches to

worldwide language sampling, where language families are represented by one or few data points (e.g. Miestamo et al. 2016), and fine-grained family-based sampling, where individual language families are investigated in greater detail (e.g. Macklin-Cordes & Round 2022).

3. Coding design for typological variables

3.1 Principles of coding design

Our approach relies on two assumptions about how languages change in contact situations. First, it assumes that contact effects depend on the intensity of contact: the more intense the interactions, the more substantial the changes. Secondly, it assumes that contact-induced changes may often start out locally, restricted to subparts of the grammar, and first may be optional before becoming obligatory. These two assumptions imply that many contact-related changes may be rather minor.

Given these assumptions, the use of existing typological databases would constitute a less ideal choice for evaluating language contact. This is because typological databases tend to code only for the most frequent feature value, thereby potentially excluding language-internal variation altogether. As a result, only the dominant and/or the most frequent value for word order, for instance, may feature in a database (e.g. Dryer & Haspelmath 2013). But high frequency may lead to a conservative effect on grammar (e.g. Bybee & Thompson 1997). For instance, pronouns are much more frequent in discourse than nouns, but their (inflectional) forms also tend to change more slowly than those of nouns. Such stable features would theoretically only change in intense contact scenarios. It is therefore conceivable that the patterns documented in most typological databases would reveal only changes in stable features, and so only reflect the outcomes of exceptionally intense contact.

In our work, we take language-internal variation as a key factor that may indicate the sources and/or direction of change in contact situations. Bickel's (2010) multivariate typology provides a feasible starting point for researching language-internal variation typologically. Once variation is captured in sets of features pertaining to different linguistic variables, we run a probabilistic assessment of clusters of properties and correlations (see §4.2). It is only after those analyses that we evaluate the extent to which contact may have shaped the structures investigated. The assessments can be flexibly adjusted to one's hypothesis; for instance, one can focus on the behaviour of linguistic features (or feature sets) by aggregating contact evidence from the language sets, or on languages by aggregating

contact evidence from the linguistic features (or their subset). The approach presented here thus marries solid coding procedures from broad typological surveys with analyses based on the observation well known to historical linguists that variation leads to change (e.g. Guy 2011). Our coding design thus presumes that contact effects will reveal themselves once pairs of languages in contact are coded for well-established variables investigated in the linguistic typology literature.

Instead of focusing on single features, say "changes to two-consonant onsets" in a given language, our method seeks to describe the behaviour of linguistic variables such as syllable structure more broadly. For this project, we are interested in five variables: (i) syllable structure, (ii) lexical prosody systems, (iii) adnominal demonstratives, (iv) nominal number, and (v) adnominal possession. Based on our experience, reference materials tend to contain detailed information on those variable domains, thereby increasing the likelihood of data representation (these also compare favourably to Lesage et al.'s 2022 review).[1]

Large typological databases such as the World Atlas of Language Structures (WALS; Dryer & Haspelmath 2013) serve as a starting point for our coding, providing a skeleton for the types of features on which we would likely find information. We thus use the operationalisation that is already available for typological descriptions of linguistic properties, such as Maddieson's (2013) criteria for analysing the syllable. We further expand on the existing coding criteria to capture potential loci of variation in different languages. For instance, in the domain of syllable structures, when collecting information about possible consonant clusters, we distinguish between consonant clusters that involve any consonant in combination with nasals (CN) and consonant clusters that involve any consonant in combination with stops (CT), among others. We do this because we assume that there might be some language-internal variation in the types of consonant clusters that may change as a result of contact. Collapsing these cluster types would potentially limit our understanding of the changes to syllable structure.

For the coding itself, we adopt four principles already in use in typological research (Witzlack-Makarevich et al. 2022):

1. Modularity and connectivity. Each of the five variables works as a stand-alone database, with every domain being coded by at least two researchers in our team. The modules are nonetheless connected through both project-internal

1. Of these five variables, lexical prosody tends to be the most underexplored, perhaps due to the fact that most reference grammars are seemingly written by morphosyntacticians and, generally speaking, with a focus on segmental phonology. Additionally, we opted to look at phonological variables other than segment inventories in a deliberate choice to increase our understanding of suprasegmental variables from a typological perspective, following Napoleão de Souza and Sinnemäki (2022).

IDs and general identification mechanisms, such as ISO codes. This principle allows for maximising the total amount of data analysed as well as introducing a level of control to the analysis itself, given that researchers in each module double-check the coding of their fellows.

2. Autotypologising. This principle mostly applies in the initial stages of the research, meaning that aspects of the coding can be adjusted when languages introduce properties for which the previously established feature values prove insufficient. Such changes take place until the coding stabilises and is deemed appropriate for the description of most languages. For instance, we observed that three-consonant onsets only ever start with /s/ in many languages. This led to the creation of a new feature assessing whether /s/-CC was the only three-consonant onset present in the language or if other combinations also occurred (e.g. /bgw/).
3. Definition files and data files. This principle gives rise to two layers of data about each possible feature: the raw data for statistical analyses (i.e. data files), as well as a qualitative description of different phenomena (i.e. definition files). Additionally, definition files provide detailed descriptions of the coding procedure, which often contain examples and reflect the current knowledge of the typology of a given variable.
4. Late aggregation. This principle means that data are collected at the lowest possible level of relevant detail, and any data aggregation takes place only after the data have been collected. Late aggregation is a central principle that makes it possible to detect even minor contact-driven linguistic changes. To improve our chances of detecting language contact that has resulted in change, data points are aggregated for as many features as possible per language in each variable domain.

In total, we have approximately 200 features corresponding to the five linguistic variables that we investigate. Variables encode roughly 25 features at a minimum. For the sake of brevity, here we only use syllable structure to illustrate our procedure. Summaries of the definition files for each of the other variables are given in Supplement S1.

3.2 Example: Syllable structure

As with the other variable domains, the coding of syllable properties expands on that used in existing typological databases. In this case, WALS served as a starting point. Maddieson (2013) codes syllables using the following criteria: (a) the number of segments in the onset, (b) the number of segments in the coda, (c) whether segments that occur as second members of two-consonant clusters are

glides and/or liquids. In our procedure, not only do we note the number of consonants in both the onset and coda, but we also perform a detailed assessment of which kinds of consonants (i.e. glides, rhotics, laterals, fricatives, etc.) occur in different combinations. Additionally, we code for the occurrence of syllabic consonants (cf. Easterday 2019), the occurrence of geminates, and further examine all types of two-consonant clusters (Easterday & Napoleão de Souza 2015). Having a more comprehensive coding such as this allows us to uncover variation that may occur in contact situations.

For instance, two-consonant onsets may enter a Focus language via borrowing, but those clusters may be restricted to /s/-C shapes (Napoleão de Souza & Sinnemäki 2022). That is, changes in structure may only be apparent in subsections of the grammar, first affecting a given feature in a mostly local fashion. Since we hypothesise that more intensive contact has a deeper influence on linguistic structure, identifying specific instances of potential diffusion could uncover pathways of contact-induced change that may become opaque as the contact evolves.

The analysis of minor patterns of structure diffusion speaks directly to the issues discussed in Matras and Sakel's (2007) classification of "matter vs. pattern" borrowing. The introduction of a syllable shape such as /st/- into a language that previously constrained onsets to being single consonants is an instance of "matter borrowing". This process has the potential to further alter the syllable structure of the language, namely its onsets, which would then constitute "pattern borrowing". For those reasons, we were especially interested in descriptions of loanword phonology, marginal structures and other so-called minor patterns in the languages in our sample.

4. Evaluating evidence for contact

4.1 Analytical grid for evaluating evidence for contact

We evaluate evidence for language contact in two steps. The first step is understanding which inferences are logically possible when comparing the Focus, Neighbour and Benchmark languages. Applying schematic binary features to the sampling triplet results in four logical outcomes in contact situations, illustrated in Table 1.

Consider a situation in which all three languages under study (Focus, Neighbour and Benchmark) share the same feature value for a given binary feature (e.g. the presence of /sn/ clusters in syllable onsets). In such cases, it is impossible to determine if this similarity between the languages is the result of inheritance, contact or universal pressures. Following Fortescue (1998), we call this type of

outcome a "mesh". Since the Focus and Neighbour languages are unrelated in our sampling, inheritance is categorically ruled out as a source for any similarity between them. In mesh situations, similarities could then derive from larger areal patterns, universal pressures, or a combination of those; this also applies to similarities between Benchmark and Neighbour languages. Because contact cannot be disentangled from either inheritance or universal tendencies in mesh situations, meshed features are excluded from further analysis.

Table 1. Logical outcomes when applying our sampling to situations of language contact. Note that the outcomes would still apply if all of the "yes" and "no" values were inverted.

Language type	Outcome			
	Mesh	**Stability**	**Convergence**	**Divergence**
Neighbour	yes	no	yes	yes
Focus	yes	yes	yes	no
Benchmark	yes	yes	no	yes

The analysis of the sample sets results in three other logical outcomes. If a feature value is the same in both the Focus and Benchmark but different in the Neighbour, this suggests stability: the feature has been inherited from the proto-language from which the Focus and Benchmark both descend ("stability" in Table 1). If the Focus and Neighbour have a feature value in common that the Benchmark does not share, this suggests that contact resulted in convergence ("convergence" in Table 1).[2] Finally, if a feature value in the Focus language differs from the values in both the Neighbour and the Benchmark, it suggests that contact may have led to innovation or divergence ("divergence" in Table 1).[3]

2. Focus and Neighbour languages may be similar to each other by sharing either the presence or the absence of a feature in comparison to the Benchmark language. While shared presence obviously provides stronger evidence for contact than shared absence, both assessments are made given the control structure in the Benchmark, which, we think, importantly contributes to their reliability (for similar considerations, see also Di Garbo and Napoleão de Souza 2023: 578).

3. As noted by a reviewer, currently our method does not distinguish independent changes (convergent evolution) from contact-induced change or from shared inheritance. For instance, it is possible that a similarity between the Focus and the Benchmark ("stability") results not from shared ancestry but from independent (parallel) changes. Likewise, it is possible that a similarity between the Focus and the Neighbour ("convergence") results not from contact but from independent non-contact-induced change. There are methods in biology for identifying independent changes (e.g. McGhee 2011), but as far as we are aware, similar computational

These four logical outcomes allow us to make inferences about change vs. stability in contact situations in a principled way. First, we remove the mesh features: this lets us focus on those features that suggest change or stability. Our focus here is on convergence vs. stability, so we also remove cases that suggest divergence.[4] Having removed the cases where the Benchmark and Neighbour share feature values, the remaining feature set functions as what we call the "baseline set" for comparing the probability of convergence to that of stability. Defined in this way, the baseline set provides a feature set in which the Neighbour and Benchmark have different feature values. In other words, the baseline includes those features where the Focus language has potentially become converged with the Neighbour language. The task is then to evaluate which of those features provide evidence for convergence, assuming that the Focus language has otherwise remained similar to the Benchmark unless contact has taken place. Here, we assume that similarities between the Focus and the Neighbour (i.e. "convergence" in Table 1) have resulted from contact, and dissimilarities (i.e. "stability" in Table 1) stem from inheritance.

Within the baseline set, the distribution of similarities and differences between the Focus and the Neighbour can be interpreted as a binomial distribution of "successes" (F = N) and "failures" (F ≠ N). Successes suggest convergence; we call those the "similarity set". The number of features in the baseline set (N_B) and in the similarity set (N_S) yield a similarity score between the Focus and the Neighbour, shown in Equation (1):

$$similarity\ score = \frac{N_S}{N_B} \tag{1}$$

The sampling scheme thus makes it possible to count the number of features that suggest evidence for convergence in a given variable domain. The similarity score produces a fraction, for instance 2/5, which estimates how strong the evidence that the Focus has converged with the Neighbour is in that domain.

This fraction provides information about the outcome of contact. In decimal form, it gives an estimate of how much contact effects dominate over inheritance: proportions above 0.5 indicate that contact effects dominate inheritance in the baseline set, while proportions below 0.5 indicate the opposite. However, simply turning the fractions into decimals is a less reliable indicator of underlying probabilities for contact effects. Even though different fractions, such as 2/5 and 8/20, are equal in decimal form (0.4), intuitively the latter would provide stronger evi-

methods have not yet been applied in linguistics. Identifying independent changes computationally requires further work on our method in future research.

4. In principle, the probability for divergence can be estimated in an analogous way if divergent feature values are not excluded.

dence about the actual impact of contact since it derives from a greater number of features. We come back to this issue in §4.2 and demonstrate how those fractions can be used for assessing evidence for contact in a Bayesian framework.

4.1.1 *Example: Syllable structure*

The grid presented in Table 1 allows for an evaluation of contact once the data have been decomposed into a number of specific features and analysed in all the sample triplets.

We then run a binary assessment of the presence of features in each of the three languages in the sampled sets: Focus, Neighbour and Benchmark. The baseline set is then identified by assigning <1>s for every dissimilarity between the Neighbour and the Benchmark languages and <0>s for every similarity between them. Next, we identify the similarity set by assigning <1>s for every similarity between the Focus and the Neighbour languages if and only if the Focus differs from the Benchmark. This analysis is illustrated with an example in Table 2 for a subset of the syllable structure domain, namely two-consonant onsets.

Table 2. Example of coding of a subset of features regarding the domain of syllable structure. C stands for any consonant; G stands for glides (e.g. /w, j/), L for laterals (e.g. /l, ʎ/); R for rhotics (e.g. /ɾ, ɽ/); N for nasals (e.g. /m, ɲ/); F for fricatives (e.g. /s, ʝ/) and T for stops (e.g. /b, q/)

Language	Type	CG	CL	CR	CN	CF	CT	Sources
South Saami	Focus	yes	yes	yes	no	no	yes	Ylikoski (2022)
Mainland Scandinavian	Neighbour	yes	yes	yes	yes	no	yes	Riad (2014)
Skolt Saami	Benchmark	yes	yes	yes	no	no	no	Miestamo (2011)
Baseline set		0	0	0	1	0	1	
Similarity set		0	0	0	0	0	1	

The Focus language in Table 2 (South Saami) shows some similarities with the Neighbour language (Mainland Scandinavian) regarding syllable onsets. However, cases in which all three languages share a feature value, as in the occurrence of CG (e.g. /kj/), fail to provide evidence of convergence in our coding, as do cases in which a structure is similar in the Focus and the Benchmark (in this case, Skolt Saami). In the example above, only one feature value provides evidence of convergence: the Focus and the Neighbour languages both have CT clusters (e.g. /sk/), which are absent in the Benchmark.

Altogether, two features differ between the Neighbour and the Benchmark, so the baseline set is two for the features pertaining to syllable onsets (CT and CN).

Out of these two features, only one feature (CT) shows evidence for convergence. Thus, the feature two-consonant onsets in South Saami yields a similarity score of 1/2.

4.2 Bayesian assessment of evidence for contact

The second step in our approach is a Bayesian assessment of evidence for contact. Our model makes two assumptions. First, wherever Neighbour and Benchmark differ in a feature (the stability or convergence cases in Table 1), we assume that the Focus originally took the same value as the Benchmark language. Second, we assume an underlying probability value p of change due to contact – termed the "change rate". If some feature in the data shows either convergence or stability between our three languages, then the probability of it being a case of convergence is the convergence rate p.

This type of model is an instance of a Bernoulli process. Data items that give us evidence for or against convergence are assumed to be independent, so all features reflect a single underlying likelihood of exhibiting convergence (the aforementioned convergence rate). The Beta distribution is a conjugate prior for Bernoulli processes. We can formally define a conjugate prior $C(h|d)$ to be a conditional probability over some parameter t, and that value $t=t(d)$ can be computed from the data, as in Equation (2).

$$\mathrm{P}(h|d_1) = C(h|t(d_1)) = \frac{P(d_1|h)}{P(d)} \quad [P(h|d_0) = C(h|t(d_0)] \tag{2}$$

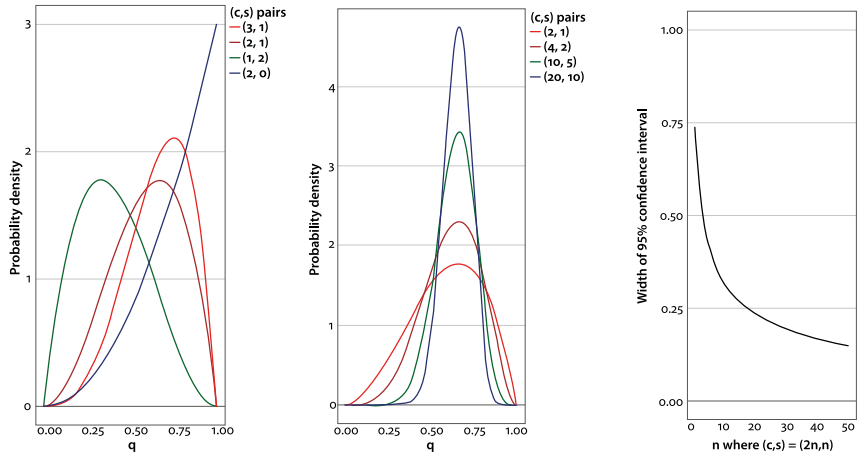
In the case of Bernoulli process data, if the prior distribution over the process parameter q is an instance of the Beta function $B=B(q;a,b)$, then the posterior is also a Beta function, albeit one with different parameters. For a Bernoulli process generating a instances of one result (e.g. "yes") and b instances of another result (e.g. "no"), the posterior distribution over the defining parameter is given by the formula in (3).

$$B(q;\, a+1,b+1) \tag{3}$$

This reasoning can be applied to evaluating convergence vs. stability with the sampling triplet. If, for a triplet (Focus, Benchmark, Neighbour), we find c instances of convergence and s instances of stability, then the distribution over possible convergence rates is given by $B(q;\, c+1,s+1)$. Figure 3a illustrates a selection of distributions arising from different pairings of convergence and stability counts. Figure 3b shows the effect of increasing the size of data, while keeping the

rate of convergence to stability the same. There is a narrowing (and consequently, a steepening) of the distribution around the value *c/(c+s)*. Where the distribution is narrower, there is less uncertainty around the value of the inferred parameter, that is, the convergence rate. Figure 3c shows how the width of the 95% confidence intervals narrows around the value 2/3 or distributions of the form *B(q; 2m+1,m+1)* for different values of *m*.[5]

This is illustrated with another example from our data on syllable structure, with the Focus language Santali (Austroasiatic), its Benchmark language Gata', and the Neighbour language Bengali with which Santali has had substantial contact (See Supplement S2 for the data). In the six features examined that showed either convergence or stability, we found four cases of convergence. Consequently, our estimate for the similarity score, our convergence parameter, is 4/6. The distribution of the likely value of the underlying preference for convergence is shown in Figure 4. The 95% confidence interval stretches from 0.29 to 0.90.



a. **b.** **c.**

Figure 3. Three graphs showing the beta distribution responding to different response counts in sampled data. Plot (a) compares distributions over the underlying probability given the counts shown. The same is true in plot (b), except that all data pairs share the same ratio of c:s being 2:1. In plot (c) we see how the 95% confidence interval narrows as the absolute counts of errors increase. Here the sample data always has a 2:1 ratio, with 3n samples in total

5. The computations were done in the R programming environment (R Core Team 2023). See Supplement S3 for the R function used for implementing the beta function.

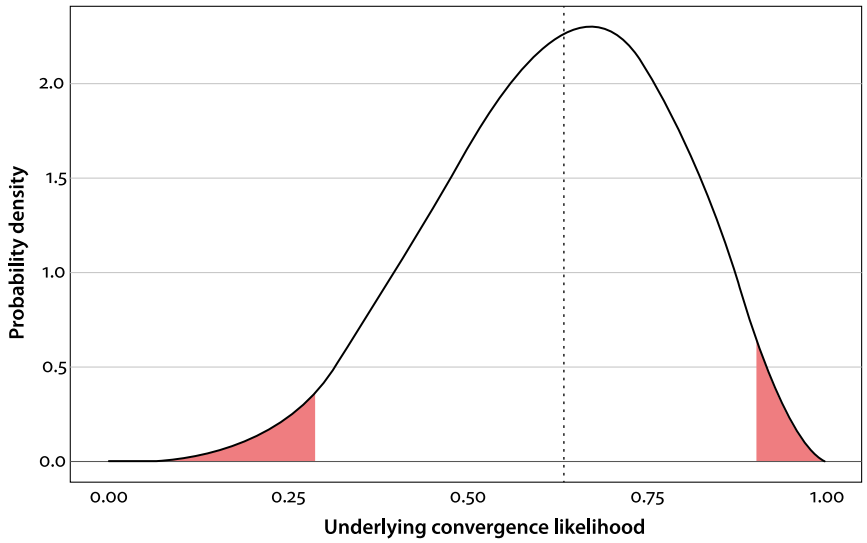


Figure 4. The distribution of the posterior likelihood of underlying levels of convergence given that we see four convergences and two stabilities in the data from the example triplet (Santali as the Focus). The best guess for this parameter – the median of this distribution – lies at 0.636, shown here with a vertical dotted line. The top and bottom 2.5% of the distribution are highlighted in red (leaving a 95% confidence interval between 0.29 and 0.90).

Inferring a distribution over how much linguistic convergence has occurred is more realistic than presuming that we can establish an exact figure. We can further work with those numbers to make more accurate assessments of aggregate quantities, using Monte Carlo simulation.

This procedure is, of course, a general principle in computing with distributions. We can numerically construct a sample of the output distribution by repeatedly selecting distributions corresponding to the input values and selecting a sample from each distribution. The more often a value appears in the outputs computed from the samples, the higher the probability of this value being correct – assuming that there is a correct value.

This Monte Carlo approach to computing aggregate quantities and relationships automatically takes into account that some individual data items might have a stronger evidential basis (and thus greater precision in their estimates) than other items. All the information in wide or narrow distributions is utilised in the combination, but the weight given to any item is exactly what it deserves given its evidential support.

In recent experiments (Napoleão de Souza et al. 2022; Sinnemäki et al. 2023), we used 27 language triplets to determine whether there is a correlation between

the probability of linguistic convergence, as estimated above, and an independently measured score of intensity of contact.[6] For each triplet, we randomly selected a single value from their distributions over the strength of language convergence. We then correlated the values as if these were exact values with our measure of the intensity of social contact. This process was repeated 10,000 times, selecting at random representatives from the probability distributions over values and computing the correlation coefficient. The resulting set of correlation coefficients constitutes a sample of the exact distribution of correlation coefficients between linguistic convergence and social contact intensity which was found in this data set. The two experiments showed that intensity of contact accounts for the probability of convergence when aggregating the linguistic variables, but not otherwise.

5. Discussion and conclusion

In this article, we have presented a typological approach for evaluating effects of language contact. We illustrated our approach with syllable structure data and demonstrated how the evidence for contact can be aggregated from low-level features. Additionally, we presented a Bayesian evaluation of the gathered evidence that takes into account the inherent uncertainty of making inferences about contact.

This method is parsimonious, as it evaluates evidence for contact in the Focus language by anchoring the inference to just two other languages, namely the Neighbour and the Benchmark. A crucial and novel component of the method is the role that the Benchmark language plays, serving as an external source of control. As mentioned above, the lack of proper control data has often been a point of criticism against contact research (e.g. Torres Cacoullos & Travis 2018). We highlighted that the method provides a starting point for making inferences about contact in a large-scale comparative way.

A further important component of the method is that it lets us assess the uncertainty in making inferences about contact. We argued that assessing this uncertainty is more realistic than presuming that an exact figure could be established for an underlying unknown probability of convergence. One suggested alternative to our method would be to compute a dissimilarity matrix directly from the feature values for each triplet (Focus, Neighbour and Benchmark). However, dissimilarity matrices produce distances between the sampling units in dec-

6. An aggregate measure of intensity of contact between Focus and Neighbour language pairs was obtained using sociolinguistic data that we collected through a sociolinguistic questionnaire. We developed this questionnaire as part of our larger project (Kashima et al., in review).

imal form and are thus not as helpful in estimating the uncertainty inherent in inferences as the method we propose here.

We have argued that the method provides reasonable estimates about convergence from a typological perspective but also acknowledge that further development is certainly possible. Even though the use of a single Benchmark language is a likely point of contention, our approach easily allows for including more Benchmarks as particular research projects require (see Supplement S4 for some ideas). A case study on contact between the Austronesian language Alorese and the Timor-Alor-Pantar language Adang suggests that inferences about contact drawn using a single Benchmark did not significantly deviate from those drawn using Bayesian ancestral state reconstruction based on a substantial portion of the family (Sinnemäki & Ahola 2023). A single Benchmark may thus be sufficient for making broad estimates of contact in worldwide samples.

We also acknowledge that the Benchmark language may have itself changed as a result of language contact or internal developments. Contact inferences on individual data points may thus be rendered false positives or false negatives due to other factors. Nonetheless, we expect that those effects would be less systematic than those we observe in the selected Focus–Neighbour contact situation. In addition, given the structure of our coding and the sheer number of features, we anticipate that the number of false positives or negatives will lose importance in the large-scale comparison. These considerations do not preclude the separate investigation of changes in the Benchmark language in future research.

Another source of possible uncertainty in our inferences lies in the fact that a Focus language may have been in contact with multiple Neighbour languages at different points in time (an issue that is particularly crucial in certain regions of the world, such as South America). The contact inferences on individual data points may thus be rendered false positives or false negatives due to contact between a Focus and some other Neighbour language(s). Where possible, we have mitigated this uncertainty by selecting a Neighbour language that has been in contact for the longest time with the Focus (often a specialist on the language family was consulted in making this choice). Still, future studies focusing on individual linguistic areas and contact scenarios could easily widen the scope of our analyses by including more than one Neighbour per Focus language. Such an expansion would provide a suitable way of assessing how multiple waves of contact may have affected structures of the Focus languages. Yet another promising avenue for development of the method would be fine-tuning estimates of the degree of convergence between languages by adding control languages for Neighbour languages as well. Initial work that adds Neighbour languages and controls for those languages is already in progress.

We have shown here how a typological analysis of descriptive sources can be used for making inferences about contact. As such, the same criticisms that befall data in typological studies may be raised regarding the current approach. Using descriptive material for making inferences about contact-induced change depends on the quality of the particular description, and also on the extent to which contact-related phenomena have been documented. Source criticism is an essential part of the research process, as in typological research. This exercise can, in itself, enable the analyst to detect instances of data pertaining to potential contact phenomena, such as borrowings, preventing them from being excluded from the analysis.

We recognise that the production of descriptive data may sometimes favour language-internal patterns to the detriment of detailed explorations of external influence (see Marten and Petzell 2016; Lüpke 2019 for criticisms of this kind of data treatment). While such methodological decisions by authors may indeed influence what inferences can be made about contact, they arguably result in underestimation rather than overstatement of contact effects.

While research in language typology usually draws on data from descriptive material, language corpora are also increasingly used (e.g. Levshina 2019). Our methodological approach can be adapted to the use of data from corpora as well, although more attention should then be paid to finding the right kinds of comparative variables and for making inferences about contact from aggregating usage-level evidence for contact. Even then, the heart of the method is to use control data in a parsimonious way.

In this article, we have argued that a typological approach to measuring language contact, in particular the probability of convergence, is not only possible but also feasible in practice. This approach enables a direct comparison of outcomes of language contact across contact settings, allowing us to move away from an overreliance on individual case studies. The method presented here can be further fine-tuned to address other kinds of problems associated with contact research, such as how we can disentangle contact from universal tendencies, and how we might incorporate the sociolinguistic context into contact research (see Supplement S4).

Funding

This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 805371; PI Kaius Sinnemäki). T. Mark Ellison was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation, Project ID 281511265, SFB 1252 "Prominence in Language") in the project C09 "Prominence and Predictive Modelling" at the University of Cologne.

Acknowledgements

An earlier version of this article was presented at the 25th International Conference on Historical Linguistics (ICHL25), 1–5 August 2022, in Oxford. We are grateful to the audience for their comments, and to two anonymous reviewers for their constructive feedback on an earlier version of this manuscript.

Statement of author contributions

The first author, Kaius Sinnemäki is the project's PI and contributed to all aspects of the work, from the study design and the implementation of the quantitative methods presented in the paper to the write-up. Francesca Di Garbo and Ricardo Napoleão de Souza contributed to the study design and the write-up. Finally, Mark Ellison contributed to the technical implementation of the quantitative methods presented in the paper and the writing of §4.2.

References

Backus, Ad. 2014. Towards a usage-based account of language change: Implications of contact linguistics for linguistic theory. In Robert Nicolaï (ed.), *Questioning language contact*, 91–118. Leiden: Brill.

Bickel, Balthasar. 2010. Capturing particulars and universals in clause linkage: A multivariate analysis. In Isabel Bril (ed.), *Clause-hierarchy and clause-linking: The syntax and pragmatics interface*, 51–102. Amsterdam: John Benjamins.

Bickel, Balthasar, Alena Witzlack-Makarevich, Kamal K. Choudhary, Matthias Schlesewsky & Ina Bornkessel-Schlesewsky. 2015. The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PLoS ONE* 10(8). e0132819.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine A. Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga & John B. Lowe. 2022. *The AUTOTYP database, version 1.1.0*. Zenodo.

Bowern, Claire. 2013. Relatedness as a factor in language contact. *Journal of Language Contact* 6(2). 411–432.

Bybee, Joan L. & Sandra A. Thompson. 1997. Three frequency effects in syntax. *Berkeley Linguistic Society (BLS)* 23. 65–85.

Cathcart, Chundra, Gerd Carling, Filip Larsson, Niklas Johansson & Erich Round. 2018. Areal pressure in grammatical evolution: An Indo-European case study. *Diachronica* 35(1). 1–34.

Croft, William A. 2021. A sociolinguistic typology for languages in contact. In Enoch O. Aboh & Cécile B. Vigouroux (eds.), *Variation rolls the dice: A worldwide collage in honour of Salikoko S. Mufwene*, 2–56. Amsterdam: John Benjamins.

Di Garbo, Francesca, Eri Kashima, Ricardo Napoleão de Souza & Kaius Sinnemäki. 2021. Concepts and methods for integrating language typology and sociolinguistics. In Silvia Ballarè & Guglielmo Inglese (eds.), *Tipologia e Sociolinguistica: verso un approccio integrato allo studio della variazione: Atti del Workshop della Società Linguistica Italiana 20 settembre 2020*, 143–176. Milan: Officinaventuno.

Di Garbo, Francesca & Ricardo Napoleão de Souza. 2023. A sampling technique for worldwide comparisons of contact scenarios. *Linguistic Typology* 27(3). 553–589.

Dryer, Matthew S. & Martin Haspelmath (eds.) 2013. *WALS Online* (v2020.3) [Data set]. Zenodo. Available online at https://wals.info

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russel D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473(7345). 79–82.

Easterday, Shelece. 2019. *Highly complex syllable structure: A typological and diachronic study*. Berlin: Language Science Press.

Easterday, Shelece & Ricardo Napoleão de Souza. 2015. Is there evidence for a hierarchy in the synchronic patterning of syllable onsets? *11th Conference of the Association for Linguistic Typology*, Albuquerque, USA, August 1–3.

Fortescue, Michael D. 1998. *Language relations across Bering Strait: Reappraising the archaeological and linguistic evidence*. London: Cassell.

Guy, Gregory R. 2011. Variation and change. In Warren Maguire & April McMahon (eds.), *Analysing variation in English*, 178–198. Cambridge: Cambridge University Press.

Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670.

Hübler, Nataliia. 2022. Phylogenetic signal and rate of evolutionary change in language structures. *Royal Society Open Science* 9(3). 211252.

Jäger, Gerhard & Johann-Mattis List. 2018. Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists. *Language Dynamics and Change* 8(1). 22–54.

Kashima, Eri, Francesca Di Garbo, Olesya Khanina & Ruth Singer. In review. The design principles of a sociolinguistic-typological questionnaire for language contact research. *Language Dynamics and Change*.

Koptjevskaja-Tamm, Maria. 2010. Linguistic typology and language contact. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 568–590. Oxford: Oxford University Press.

Lesage, Jakob, Hannah J. Haynie, Hedvig Skirgård, Tobias Weber & Alena Witzlack-Makarevich. 2022. Overlooked data in typological databases: What Grambank teaches us about gaps in grammars. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2884–2890. Marseille: European Language Resources Association. https://aclanthology.org/2022.lrec-1.309

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

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.

Lüpke, Friederike. 2019. Language endangerment and language documentation in Africa. In H. Ekkehard Wolff (ed.), *The Cambridge handbook of African linguistics*, 468–490. Cambridge: Cambridge University Press.

Macklin-Cordes, Jayden L. & Erich R. Round. 2022. Challenges of sampling and how phylogenetic comparative methods help: With a case study of the Pama-Nyungan laminal contrast. *Linguistic Typology* 26(3). 533–572.

Maddieson, Ian. 2013. Syllable structure. In Matthew S. Dryer & Martin Haspelmath (eds.), *WALS Online* (v2020.3) [Data set]. Zenodo. Available online at http://wals.info/chapter/12

Marten, Lutz & Malin Petzell. 2016. Linguistic variation and the dynamics of language documentation: Editing in 'pure' Kagulu. *Language Documentation & Conservation* 10. 105–129. http://hdl.handle.net/10125/24651

Maslova, Elena. 2003. A case for implicational universals. *Linguistic Typology* 7(1). 101–118.

Matras, Yaron & Jeanette Sakel. 2007. Investigating the mechanisms of pattern replication in language convergence. *Studies in Language* 31(4). 829–865.

McGhee, George R. 2011. *Convergent evolution: Limited forms most beautiful*. Cambridge, MA: MIT Press.

Miestamo, Matti. 2011. Skolt Saami: A typological profile. *Suomalais-Ugrilaisen Seuran Aikakauskirja* 2011(93). 111–145.

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

Napoleão de Souza, Ricardo, Francesca Di Garbo, Kaius Sinnemäki, Eri Kashima, Noora Ahola, Anu Hyvönen & Oona Raatikainen. 2022. Typologizing contact effects on a global scale. Paper presented at the *14th Biennial Conference of the Association for Linguistic Typology*, 15–17 December 2022, Austin, TX.

Napoleão de Souza, Ricardo & Kaius Sinnemäki. 2022. Beyond segment inventories: Phonological complexity and suprasegmental variables in contact situations. *Journal of Language Contact* 15(3–4). 439–480.

Neureiter, Nico, Peter Ranacher, Nour Efrat-Kowalsky, Gereon A. Kaiping, Robert Weibel, Paul Widmer & Remco R. Bouckaert. 2022. Detecting contact in language trees: A Bayesian phylogenetic model with horizontal transfer. *Humanities and Social Sciences Communications* 9(1). 1–14.

Polinsky, Maria. 2014. Heritage languages and their speakers: Looking ahead. In Marta Fairclough & Sara M. Beaudrie (eds.), *Innovative approaches to heritage languages: From research to practice*, 325–346. Washington, DC: Georgetown University Press. https://dash.harvard.edu/handle/1/33946918

R Core Team. 2023. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

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

Riad, Tomas. 2014. *The phonology of Swedish*. Oxford: Oxford University Press.

Sinnemäki, Kaius, Francesca Di Garbo, Eri Kashima, Ricardo Napoleão de Souza & T. Mark Ellison. 2023. Language contact effects in their multilingual ecology: A typological approach. A paper presented at the *56th Annual Conference of the Societas Linguistica Europaea (SLE)*, 29 August–1 September 2023, Athens.

Sinnemäki, Kaius & Noora Ahola. 2023. Testing inferences about language contact on morphosyntax: A typological case study on Alorese–Adang contact. *Transactions of the Philological Society* 121(3). 513–545.

Thomason, Sarah Grey. 2001. *Language contact: An introduction*. Edinburgh: Edinburgh University Press.

Thomason, Sarah Grey & Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Torres Cacoullos, Rena & Catherine E. Travis. 2018. *Bilingualism in the community: Code-switching and grammars in contact*. Cambridge: Cambridge University Press.

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

Witzlack-Makarevich, Alena, Johanna Nichols, Kristine A. Hildebrandt, Taras Zakharko & Balthasar Bickel. 2022. Managing AUTOTYP data: Design principles and implementation. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.), *The open handbook of linguistic data management*, 631–642. Cambridge, MA: MIT Press.

Yakpo, Kofi. 2020. Social factors. In Evangelia Adamou & Yaron Matras (eds.), *The Routledge handbook of language contact*, 129–146. London: Routledge.

Ylikoski, Jussi. 2022. South Saami. In Marianne Bakró-Nagy, Johanna Laakso & Elena Skribnik (eds.), *The Oxford guide to the Uralic languages*, 113–129. Oxford: Oxford University Press.

Supplements

Supplement S1. Descriptions of the broad typological categories

Summary of the definition files

In this section, we provide short summaries of the definition files we use for the four linguistic variables beyond syllable structure, that is, lexical prosody, nominal number, adnominal possession, and demonstrative systems. Here we only highlight the general design principles that guided our linguistic data collection given that, in this article, the focus is on methodology rather than on the empirical analyses stemming from this coding procedure.

Nominal number

Number is possibly the most frequent of all nominal categories (Corbett 2000; Igartua 2015), that is, likely to be expressed in some form across all language families and continents. In addition, cross-linguistically, number systems span the lexicon, morphology, and syntax, which means that potential contact effects in number systems can occur in a variety of domains of language structure.

Number is a morphosyntactic category which encodes quantification in relation to ENTITIES, denoted by nominals, or EVENTS, denoted by verbs (Kibort & Corbett 2008). Number systems presuppose that the possibility to construe something as a token (of an entity or an event) and to differentiate between one and more than one instance of that token is grammaticalized to a certain degree. For instance, in English, the opposition between one or more instances of the token 'cat', is coded by the opposition between the forms *cat* and *cats*. Similarly, in Rapa Nui (Oceanic), the opposition between one or more instances of the event 'dive' is coded by the opposition between the forms *ruku* 'dive (once)' and *ruku ruku* 'dive repeatedly/go diving'.

The focus of our project is on NOMINAL NUMBER, that is, the wholesale strategies that languages use for the quantification of entities, which we call NOMINAL NUMBER SYSTEMS (or simply number systems).

Our coding design is based on earlier work by Kibort & Corbett (2008) and targets the following dimensions of number systems:

1. the NUMBER VALUES distinguished in a language (e.g., singular, plural, dual, trial)
2. the LOCUS OF NUMBER MARKING of each number value (for instance whether nouns, pronouns, adnominal modifiers, or verbs, exhibit a singular, plural dual or trial distinction). This dimension allows us to capture whether, for each number value, number marking occurs at the noun phrase and/or at the clausal level.
3. the TYPE OF NUMBER MARKING, which considers whether number distinctions are encoded by number words (e.g., via plural words), morphologically (e.g., via affixation, stem alternation, clitics), lexically (via suppletion), or through a combination of these.
4. The OBLIGATORINESS of number marking. In several languages of the world, not all nouns participate in the encoding of number distinctions in equal terms, and number marking is not always obligatory (on nouns or elsewhere). This dimension allows us to capture which properties of the possessive systems tend to be optional and/or obligatory.
5. INTERACTIONS with the encoding of other morphosyntactic features, such as case, gender, definiteness, and person. This dimension allows us to capture aspects of these interactions, such as, for instance, cumulative exponence and patterns of syncretism with any of the above-mentioned features.

All in all, the coding scheme for nominal number encompasses 56 distinct features allowing for binary yes/no answers. A comment slot is added to each of these features in order to incorporate additional relevant information, such as the bibliographic sources which our coding decision is based upon or any relevant comments.

Adnominal possession

Adnominal possession refers to syntactic noun phrases whose head is a possessum and that may be modified by another nominal functioning as a possessor. These constructions typically express ownership, such as *my car*, part-whole relationships, such as *the leg of the table*, and

kinship relationships, such as *Lisa's daughter*. While these constructions can be used for other functions in different languages (e.g., Koptjevskaja-Tamm, 2003; Haspelmath, 2017; Ortmann, 2018), here we focus on their typical functions. Adnominal possession occurs in almost every language. Culturally, social contact that involves different dialects or languages may further highlight the importance of ownership (e.g., land ownership; Aikhenvald 2013).

The coding design is inspired by earlier typological work on possessive noun phrases by Nichols (1992), Nichols & Bickel's (2013) coding in the *WALS*, and especially by the coding for adnominal possession in the *AUTOTYP* database (Bickel et al. 2022). The analysis targets the following dimensions of adnominal possession:

1. Adnominal possessive constructions are classified according to LOCUS OF MARKING (Nichols 1992). This refers to the location where syntactic relations (dependencies) are morphologically marked in the construction: on the head of the construction (head marking), on the dependent (dependent marking), on both (double marking), or on neither (zero marking). Two rare additional types are distinguished.
2. CONTEXT OF VARIATION refers to language-internal variation in locus of marking being conditioned by some property of the possessor or the possessum. Most typically this refers to alienability distinctions, such as when, for instance, zero marking is used for inalienable possession and dependent marking for alienable possession (see Haspelmath 2017). We code for alienability distinctions, pronominal vs full noun possessors, animacy, definiteness, and person/number.
3. BOUNDNESS refers to the way morphological marking is achieved. The dependency relation between the possessor and the possessum may be morphologically marked via independent morphemes, such as adpositions, particles, and clitics, via bound means, such as affixes and morphophonological and tonal alternations, or via a combination of these.
4. LINEAR ORDER refers to the relative order of the head and the dependent. This feature is analysed with four values: head-dependent, dependent-head, both, and inapplicable (head marking constructions where the possessor is not expressed as a separate constituent).
5. The HOST-MARKER ORDER refers to the order between the morphological marker and its host. This dimension allows us to capture mainly whether the morphological marker occurs before the host noun (e.g., prefixes) or after it (e.g., postpositions).
6. The OBLIGATORINESS of morphological marking. In many languages of the world, not all possessive constructions are encoded in equal terms, and morphological marking is not always obligatory. This dimension allows us to capture which properties of number systems tend to be optional and/or obligatory.

In the analysis the constructions are first analysed in terms of locus of marking and the other features are then analysed with respect to these constructions. The reason for this is that the other features often depend on the language-internal variation in locus of marking.

The analysis is delimited to unmodified noun phrases. Expressions, such as *my new car*, are excluded, since in some languages they may behave differently from unmodified adnominal possession. Predicate possessive expressions, including predicative possessive pronouns (e.g., *mine, yours, theirs*) are further excluded. We also do not code for external possession. In Example (1) from German, the dative possessor *Mir* is not part of the same constituent with the possessum *Hände*. Moreover, the verb *zittern* is intransitive, its subject is *die Hände*, and the dative possessor *Mir* is not part of the verb's argument structure.

(1) German (Germanic, Indo-European; König 2001: 970)
Mir zittern die Hände
me.DAT shake.PL the hand.PL
'My hands are shaking.'

The coding scheme is designed to allow only for binary yes/no answers. A comment slot is added to each feature to incorporate in prose any additional relevant information, as well as the bibliographic sources which our coding decision is based upon.

Lexical prosody systems

Lexical prosody refers to the lexical prominence (lexical stress) and lexical intonation (e.g., lexical tones) patterns of languages. The focus of our project is LEXICAL PROMINENCE, which often refers to lexical stress. It implies that one unit, usually the syllable, is more prominent than others within the lexical word. Most languages of the world have lexical stress (Roettger & Gordon 2017), but many also have lexical tone, including restricted tone, also known as 'pitch accent'. Another broad type of prominence type, edge-prominence (Jun 2014), will not be coded in detail, as it is still poorly understood in the field. Our coding is based primarily on Hyman (2006) and Gordon (2016).

The present coding does not describe *Intonation or Phrase-level prosody*. Those two levels of prosody also make use of prominence. However, they mostly refer to manipulations of f0 that change the function of the phrase or utterance (e.g., declarative vs interrogative), but also manipulations that convey speaker's attitude (e.g., irony, disbelief) or emotions (e.g., anger). Our coding only contemplates the (phonological) word level.

The following are the dimensions of lexical prosody systems for which we code:

1. CONTRASTIVENESS. This criterion relates to the function of prominence in a language; assessing whether phonetically similar/identical words form minimal pairs (lexical contrast) based on a suprasegmental property. It may also refer to high tones in privative tone systems.
2. LOCATION. This criterion distinguishes languages in which prominence is fixed on syllable (e.g., word-final stress) from those where it occurs in different locations. Note that while fixed prominence is never contrastive, moveable prominence may not in itself create lexical contrasts. This may be seen in languages in which, for instance, nouns have a high tone on the first syllable whereas adjectives have a high tone on the penultimate. The location window is also specified.
3. TYPE OF PROMINENCE. This criterion distinguishes primarily languages with stress from languages with lexical tone, among which we include privative tone systems and restricted tone systems. For a language to have lexical stress, it must be (a) invariable for the same words; (b) cue prominence on a single syllable/mora; (c) obligatory (cf. Hyman 2006). Note that authors speaking (Indo)-European languages often state a language has stress even if it fails to conform to the criteria proposed. Our coding may thus disagree with the reference grammar author. Following Hyman (2006), we consider 'pitch accent' languages tone languages, albeit with restrictive tone.
4. DETERMINING FACTORS. This criterion seeks to establish whether prominence derives from lexical, morphological, and/or phonological factors. When prominence depends on given affixes or word classes, it relates to morphological factors. Syllable weight is a common phonological factor determining prominence. If no clear patterns are discernible, prominence is considered lexically determined.

5. PHONETIC EXPRESSION. This criterion describes how prominence is phonetically cued in the languages, for instance, via duration in stressed syllables, the number and shapes of tones, accompanying laryngeal modifications (e.g., creaky voice), and so forth.

Comment slots allow coders to incorporate any additional information, as well as relevant references or disagreements. Due to the generalised lack of description of lexical prosody systems, we often resorted to databases such as StressTyp (Goedemans & van der Hulst 2009; Goedemans, Heinz & van der Hulst 2015) and LAPSyD (Maddieson et al. 2014–2016).

Adnominal demonstratives

Demonstratives are deictic expressions primarily used to direct the attention of participants in the speech situation to specific objects and/or locations, often in combination with a pointing gesture (Diessel 1999a, 1999b). Our focus here is on adnominal demonstratives, as well as on reinforcing deictic particles.

Demonstratives typically encode the perceived location of the entities identified in the speech situation. In addition, they often index grammatical information pertaining to the nominal referent, such as number, gender, and case. In this project we are interested in both of these aspects. More specifically, our coding design targets the following dimensions:

1. The TYPE AND NUMBER OF DISTINCTIONS IN DEMONSTRATIVE SYSTEMS: Demonstrative systems may be structured in reference to relative locations at the time of the utterance, the position of speech participants, the visibility of the referent, and/or information related to the geography of the area where a given speech act occurs. We assess whether the languages of our sample make any of these distinctions, how many values they feature for each of these distinctions, and whether the expression of these values is obligatory.
2. The MORPHOSYNTACTIC PROPERTIES OF DEMONSTRATIVES: Demonstrative systems may inflect for number, gender, and case in agreement with the noun. We assess whether any of these morphosyntactic features is expressed in the demonstrative systems of the sampled languages, and how many distinctions are made per feature.
3. BOUNDEDNESS: This dimension addresses the type of morphosyntactic encoding of demonstratives in a given language, and, in particular, whether demonstratives are free or bound morphemes.
4. ALLOMORPHY: This dimension captures whether demonstrative stems and affixes change depending on the type of spatial distinctions (e.g., proximity) and/or the morphosyntactic category (number, gender, case) that they express.
5. LINEAR ORDER: This criterion targets demonstratives that are encoded by free morphemes (separate words). It asks where demonstrative words are placed with respect to the referential expression, and whether this position changes across types of demonstrative distinctions.

A comment slot is added to each feature in order to include any additional information, such as bibliographic sources and/or coding issues.

References

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine A. Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga & John B. Lowe. 2022. The AUTOTYP database, version 1.1.0. Zenodo.

Corbett, Greville G. 2000. *Number*. Cambridge: Cambridge University Press.

Diessel, Holger. 1999a. *Demonstratives: Form, function and grammaticalization*. Amsterdam: John Benjamins.

Diessel, Holger. 1999b. The morphosyntax of demonstratives in synchrony and diachrony. *Linguistic Typology* 3(1). 1–50.

Goedemans, Rob & Harry van der Hulst. 2009. StressTyp: A database for word accentual patterns in the world's languages. In Martin Everaert, Simon Musgrave & Alexis Dimitriadis (eds.), *The use of databases in cross-linguistic studies*, 235–282. Berlin: Mouton de Gruyter.

Goedemans, Rob, Jeffrey Heinz & Harry van der Hulst. 2015. StressTyp2, version 1. http://st2.ullet.net

Gordon, Matthew. 2016. *Phonological typology*. Oxford: Oxford University Press.

Haspelmath, Martin. 2017. Explaining alienability contrasts in adpossessive constructions: Predictability vs. iconicity. *Zeitschrift für Sprachwissenschaft* 36(2). 193–231.

Hyman, Larry M. 2006. Word-prosodic typology. *Phonology* 23(2). 225–257.

Igartua, Iván. 2015. From cumulative to separative exponence in inflection: Reversing the morphological cycle. *Language* 91(3). 676–722.

Jun, Sun-Ah. 2014. Prosodic typology: by prominence type, word prosody, and macrorhythm. In Sun-Ah Jun (ed.), *Prosodic typology II: The phonology of intonation and phrasing*, 520–539. Oxford: Oxford University Press.

Kibort, Anna & Greville G. Corbett. 2008. *Grammatical features inventory: Number*. University of Surrey.

Koptjevskaja-Tamm, Maria. 2003. Possessive noun phrases in the languages of Europe. In Frans Planck (ed.), *Noun phrase structure in the languages of Europe*, 621–722. Berlin: Mouton de Gruyter.

König, Ekkehard. 2001. Internal and external possessors. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals, Volume 2*, 970–978. Berlin: De Gruyter Mouton.

Maddieson, Ian, Sebastien Flavier, Edigio Marsico & Francois Pellegrino. 2014–2016. LAPSyD: Lyon-Albuquerque Phonological Systems Databases, Version 1.0. https://lapsyd.huma-num.fr/lapsyd/

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

Nichols, Johanna & Balthasar Bickel. 2013. Locus of marking in possessive noun phrases. In Matthew Dryer & Martin Haspelmath (eds.), *WALS Online* (v2020.3) [Data set]. Zenodo . Available online at http://wals.info/chapter/24

Ortmann, Albert. 2018. Connecting the typology and semantics of nominal possession: alienability splits and the morphology-semantics interface. *Morphology* 28(1). 99–144.

Roettger, Timo & Matthew Gordon. 2017. Methodological issues in the study of word stress correlates. *Linguistics Vanguard* 3(1). 20170006.

Supplement S2. Example data from 2 sample sets for syllable structure

Token_ID	syll-27_1	syll-27_2	syll-27_3	syll-27_4	syll-27_5	syll-33_1	syll-33_2	syll-33_3	syll-33_4	syll-33_5
Language	Santali	Bengali	Gata'			Muak Sa-Aak	Lü/Tai Lue	Pnar		
Language type	1- Focus	2- Neighbor	3- Benchmark			1- Focus	2- Neighbor	3- Benchmark		
ISO	sat	ben	gaq			tlq	khb	pbv		
Glottocode	sant1410	beng1280	gata1239			tail1246	luuu1242	pnar1238		
Family	Austroasiatic	Indo-European	Austroasiatic			Austroasiatic	Tai-Kadai	Austroasiatic		
GA_set	27	27	27			33	33	33		
Area	Indic I	Indic I	Indic I			Southeast Asia I	Southeast Asia I	Southeast Asia I		
Macroarea	Eurasia	Eurasia	Eurasia			Eurasia	Eurasia	Eurasia		
References	Neukom (2001); Ghosh (2008); LAPSyD	David (2015); Mukherjee (2011)	Anderson (2008)			Hall (2010)	Hall (2010); LAPSyD	Ring (2015)		
0- Set				Similarity set	Baseline set				Similarity set	Baseline set
0a- Counts				4	6				6	7
1- Max syll shape	CVCC	CCCVC	CCVC	0	1	CCVC	CCVC	CCVC	0	0
1a- Specify 1	The only CC clusters are homorganic nasal + C (LAPSyD). In some dialects, CVNC > CṼC as in /poɳɖ/ ~ /põɖ/. Nasality is a major marker of dialect (Ghost 2008: 17).	David (2015: 24). Coda clusters only occur in the speech of highly educated persons (p. 23).	Anderson 2008: 634–635			Muak is mono- and sesquisyllabic; there are words where presyllables precede the onset forming clusters C.CC (Hall 2010: 55, 56).	CCVC, extrapolated from examples (Hall 2010: 32). Clusters given in (Hall 2010: 31). LAPDSyD categorizes as CVC because CCV occurs only in literary genres and loanwords.	Ring (2015: 31). Ring (2015: 21–22) also describes that Pnar has complex onset C-clusters which do not follow the typical sonority, and thus Pnar differs from other similar languages typologically. CCC clusters with /j/ as the third element occur (e.g. /snjoʔ/; Ring 2015: 44).		
2- CCV	no	yes	yes			yes	yes	yes		
2a- CGV	no	no	no	0	0	yes	yes	no	1	1
2b- CLV	no	yes	yes	0	0	no	yes	yes	0	0
2c- CRV	no	yes	yes	0	0	no	yes	yes	0	0
2d- CNV	no	no	yes	1	1	no	no	yes	1	1
2e- CFV	no	no	yes	1	1	no	no	yes	1	1
2f- CTV	no	no	yes	1	1	no	no	yes	1	1
2g. Specify (2a–f)		FL & FR clusters also	Different combinations			C1 can be any voiceless stop	L as C2 restricted	Ring (2015: 49)		

Supplement S2. *(continued)*

Token_ID	syll-27_1	syll-27_2	syll-27_3	syll-27_4	syll-27_5	syll-33_1	syll-33_2	syll-33_3	syll-33_4	syll-33_5
		occur (David 2015: 23).	of FC, NC and a more restricted set of RC also occur (Anderson 2008: 684).			except /c/ or unaspirated /t, d/ + C2 is a glide (/r/, /w/ or /j/) (Hall 2010: 42, 55). Author counts /r/ as a glide in this context (Hall 2010: 41).	only to one dialect and occurs very rarely, also /tw/ and /tʰw/ occur only in a dialect (Hall 2010: 31).	lists consonant clusters that are attested in onset. There seem to be quite a few CG occurreces as 'allophones' of given diphthongs, also table on p. 48.		
3- CCCV	no	yes	no	0	1	no	no	no	0	0
3a- /s/CC only	no	yes	no			NA	NA	NA		
3b- Specify 3–3a		Only /s/TR occurs (David 2015: 23).				NA	NA	NA		
4- CCCCV+	no	no	no	0	0	no	no	no	0	0
5- Syllabic C	yes	no	no	0	0	yes	no	yes	0	1
5a- Specify 5	Nasals occur as the nuclei of syllables (Ghosh 2008: 30).					Bilabial nasals can occur as syllabic but this is very rare (Hall 2010: 39). Additionally language has the C presyllables.	Syllabicity not mentioned.	According to Ring (2015: 53), a nasal, a trill, or a lateral can occur in a nucleic position.		
6- VC	yes	yes	yes			yes	yes	yes		
6a- VN	yes	yes	yes	0	0	yes	yes	yes	0	0
6b- VLq	yes	yes	yes	0	0	yes	yes	yes	0	0
6c- VF	yes	yes	no	1	1	no	no	yes	1	1
6d- VT	yes	yes	yes	0	0	yes	yes	yes	0	0
Specify (6a–d)	VF sequences only found in loanwords from Indo-Aryan (Ghosh 2008: 30).	Extrapolated from examples in David (2015).	Coda examples extrapolated from Anderson (2008); no examples of VF were found even though the language has /s h/.			For stops, only unreleased voiceless ones can occur in coda. Glide/rhotic /r/ is not allowed (Hall 2010: 43). Only two Fs in Muak /s h/, whereas /f/ is marginal.	Hall (2010: 33).	Extrapolated from examples (Ring 2015: 35). It is also stated (Ring 2015: 31, 35) that there is a general tendency for final C's to be unreleased, so the fricative /h/ does not occur in the		

Supplement S2. *(continued)*

Token_ID	syll-27_1	syll-27_2	syll-27_3	syll-27_4	syll-27_5	syll-33_1	syll-33_2	syll-33_3	syll-33_4	syll-33_5
								coda, and the affricates /tʃ/ and /dʒ/ are neutralized to /ʒ/. /l/ and /s/ occur only in the codas of borrowed words. In addition final consonant voicing is inconsistent, but is present particularly in some loan words.		
7- VCC	yes	no	no	0	0	no	no	no	0	0
7a-VLqC	no	no	no	0	0	no	no	no	0	0
7b-VNC	yes	no	no	0	0	no	no	no	0	0
7c-VFC	no	no	no	0	0	no	no	no	0	0
7d-V/s/C only	no	no	no	0	0	no	no	no	0	0
7e-VTC	no	no	no	0	0	no	no	no	0	0
7f- Specify (7a–e)	See note in max syll shape.									
8- VCCC	no	no	no	0	0	no	no	no	0	0
9- VCCCC+	no	no	no	0	0	no	no	no	0	0
9a. Specify (8–9)										
10. Geminates	no	yes	yes	0	0	no	no	yes	1	1
10a. Specify 10						C-cluster (Hall 2010: 41). No mention of geminates explictly.		Ring (2015: 46–47)		
11. LAPSyD link	https://lapsyd.huma-num.fr/lapsyd/index.php?data=view&code=10029	https://lapsyd.huma-num.fr/lapsyd/index.php?data=view&code=322					https://lapsyd.huma-num.fr/lapsyd/index.php?data=view&code=461			
12. Notes						Hall (2010) has a section on contact between Muak Sa-Aak and Lü.				

References

Anderson, Gregory D. S. 2008. Gtaʔ. In Gregory D. S. Anderson (ed.), *The Munda languages*, 682–763. London: Routledge.

David, Anne Boyle. 2015. *Descriptive grammar of Bangla*. Berlin: Mouton de Gruyter.

Ghosh, Arun. 2008. Santali. In Gregory D. S. Anderson (ed.), *The Munda languages*, 11–98. London: Routledge.

Hall, Elizabeth. 2010. A phonology of Muak Sa-aak. MA thesis, Payap University, Chiang Mai.

Maddieson, Ian, Sébastien Flavier, Edigio Marsico & François Pellegrino. 2014–2016. LAPSyD: Lyon-Albuquerque Phonological Systems Databases, Version 1.0. http://www.lapsyd.ddl.ish-lyon.cnrs.fr/lapsyd/

Neukom, Lukas. 2001. *Santali*. Languages of the World/Materials, 323. München: Lincom.

Ring, Hiram. 2015. A grammar of Pnar. PhD dissertation, Nanyang Technological University.

Sibasis Mukherjee. 2011. Bengali. In *West Bengal: Part-I*, 48–102. India: Language Division Office of the Registrar General & Census Commissioner.

Supplement S3. R function for beta distribution

```
## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ##
## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ##
## betaDistro
## Author: T. Mark Ellison
## Date of last update: 24 October 2023
## This function determines the distributions over parameters given
##    the proportions of evidence weighting in favour of one item or another.
## Parameters:
##    n1s = number of "1" valued items occurring (the similarity set)
##    outOf = number of items overall (the baseline set)
## Return value is a list with the following elements:
##    $ciBottom, $ciTop – the bottom and top of the 95% confidence intervals
##    $quartile1,$median,$quartile3 – the top values of the
##       first, second and third quartiles respectively
##    $sampler(n) – a function for producing n random samples from
##       the distribution
##    $graph – a ggplot graph object for plotting the distribution – you
##       can add more ggplot2 features to the graph if you like
## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ## ##
require(tidyverse)
require(ggplot2)
betaDistro <- function(n1s,outOf) {
  n0s <- outOf – n1s
  list(
    ciBottom = qbeta(0.025,n1s+1,n0s+1),
    quartile1 = qbeta(0.75,n1s+1,n0s+1),
    median = qbeta(0.5,n1s+1,n0s+1),
```

```
        quartile3 = qbeta(0.75,n1s+1,nos+1),
        ciTop = qbeta(0.975,n1s+1,nos+1),
        sampler = function(n) {rbeta(n,n1s+1,nos+1)},
        graph = (function(n) {
          ((0:n)/n) %>%
            data.frame(x = .) %>%
            mutate(y = dbeta(x,n1s+1,nos+1)) %>%
            (function(d) {
              ggplot(data=d) +
                geom_line(aes(x=x,y=y))
            })
        })(10000)
    )
}
```


Supplement S4. Potential improvements to the inference model

The model developed in this article enables us to estimate contact effects across contact situations by controlling for the most obvious confounding factor, namely genealogical relatedness of the languages in contact. A further confounding factor is universal preference (e.g., Ranacher et al. 2021). The Focus language may have acquired a particular new feature value because it is universally preferred across languages and not because of contact.

One way in which the current model could be further improved is by addressing universal preference using data in typological distributions. This can be done by introducing a probability that reflects the a priori likelihood of individual feature values. These likelihoods can be used as prior probabilities of particular feature values occurring (see, e.g., Ranacher et al. 2021). We envisage using such likelihoods in the following way. Suppose the *a priori* likelihood of feature F taking value v is F_v. For a specific contact triple t, we could fit two parameters $ß_t$ and ν_t, which determine how much the likelihood is shifted away from the global typological likelihood in favour of the values found in the Benchmark and Neighbour languages, presumably due to shared inheritance and contact respectively. We might then look at Bayes' Factor ratios between best fitting models from the four model classes: where $ß_t = 0 = \nu_t$, where $ß_t > 0$, $\nu_t = 0$, where $ß_t = 0$, $\nu_t > 1$, and where $ß_t > 0 < \nu_t$.

Differential probabilities of transmission of typological features – conditioned by whether Ancestor and Neighbour agree on those features – constitute evidence for a contact effect.

A further improvement in the evaluation can be achieved if we can reconstruct the most recent ancestor of the Focus language which did not yet have contact with the Neighbour language in its history. Let's call this language *Ancestor language*. It may be the most recent common ancestor with the Benchmark language, or it may be a descendant of that language as shown in Figure 5.

If there is an effect of contact, we should see a difference in the transmission of features from Ancestor to Focus when the Ancestor and Neighbour shared values for a typological feature, and when they did not. We expect the transmission of typological features to be more faithful in the former case, and less faithful in the latter case.

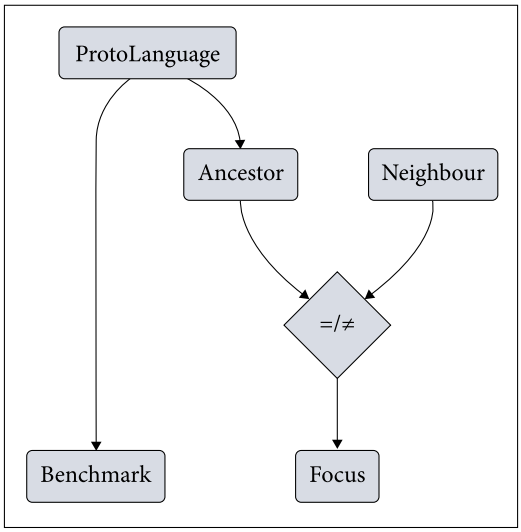


Figure 5. A diagram of an enriched history of the contact situation. Ancestor language is the most recent parent of the Focus language which had not had contact with Neighbour

We have a better chance of reconstructing the Ancestor language when several Benchmark languages are available, meaning many languages that are related to Focus but that have not been in contact with Neighbour. With data on such languages, we could reconstruct the family tree, or use an existing family tree, to infer the likely typological patterns in earlier stages of the Focus language (see Sinnemäki & Ahola 2023, for an example). This is another illustration of how our typological approach can be further tuned.

References

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

Sinnemäki, Kaius & Noora Ahola. 2023. Testing inferences about language contact on morphosyntax: A typological case study on Alorese – Adang contact. *Transactions of the Philological Society* 121(3). 513–545.

Résumé

Les phénomènes de contact sont de plus en plus étudiés à partir de différentes perspectives telles que la linguistique historique, la sociolinguistique, la linguistique typologique et aréale. Cependant, vu que la majorité de ces recherches se basent sur des études de cas, il n'y a pas, à ce jour, une évaluation systématique des phénomènes de contact à partir d'une approche comparative à large échelle. En s'inspirant de la linguistique historique et typologique, cet article présente une nouvelle approche typologique pour évaluer l'incidence du contact sur la

distribution de certaines propriétés structurelles des langues. La méthode se compose de trois parties: (1) une nouvelle approche à l'échantillonnage typologique, (2) une nouvelle procédure d'analyse pour les données typologiques et (3) une nouvelle technique statistique pour formuler des inférences sur la probabilité du changement dû au contact. Nous démontrons que cette méthode parcimonieuse, qui permet d'évaluer les effets du contact sur la structure des langues, offre un point de départ solide pour le développement des approches typologiques portant sur le contact langagier.

Zusammenfassung

Sprachkontaktphänomene werden zunehmend aus verschiedenen Perspektiven untersucht, z. B. aus historischer, soziolinguistischer oder areal-typologischer Perspektive. Allerdings basiert die Mehrheit der Untersuchungen auf Fallstudien, sodass eine Auswertung der Kontaktphänomene aus einer weltweiten komparativen Perspektive in der Literatur bis anhin fehlt. In diesem Artikel präsentieren wir einen neuen typologischen Ansatz für die Analyse von Sprachkontakt, der Inspiration aus der historischen Linguistik und Typologie zieht. Diese Analysemethode kann verwendet werden um zu evaluieren, ob gegebene linguistische Domänen von Sprachkontakt beeinflusst wurden. Unsere Methode besteht aus drei Teilen: (1) einer neuen Samplingmethode, (2) der Analyse typologischer Daten, und (3) den probabilistischen Schlussfolgerungen über Sprachkontakt. Es wird hier argumentiert, dass dies eine sparsame Methode zur Bewertung von Sprachkontakteffekten ist, und dass die Methode auch als Ausgangspunkt für Weiterentwicklungen typologischer Ansätze zu Sprachkontakt benutzt werden kann.

Address for correspondence

Kaius Sinnemäki
General Linguistics
University of Helsinki
P.O. box 24
Unioninkatu 40
FIN-00014
Finland

kaius.sinnemaki@helsinki.fi
https://orcid.org/0000-0002-6972-5216

Co-author information

Francesca Di Garbo
CNRS Laboratoire Parole et Langage (UMR 7309)
Aix-Marseille University
University of Helsinki
francesca.DI-GARBO@univ-amu.fr
https://orcid.org/0000-0002-2499-8800

Ricardo Napoleão de Souza
Linguistics and English Language
University of Edinburgh
University of Helsinki
r.n.deSouza@ed.ac.uk

T. Mark Ellison
Institut für Linguistik
University of Cologne
Universität zu Köln
t.m.ellison@uni-koeln.de

Publication history

Date received: 24 April 2023
Date accepted: 6 December 2023
Published online: 25 June 2024