Linguistic Typology at the Crossroads 5-2 (2025): 1-35 ISSN 2785-0943

Towards greater social anchoring in language typology

FRANCESCA DI GARBO[1], KAIUS SINNEMÄKI[2] , ERI KASHIMA[2,3]
[1]AIX-MARSEILLE UNIV.,CNRS LPL, [2]UNIVERSITY OF HELSINKI, [3]AUSTRALIAN NATIONAL UNIVERSITY

Submitted: 23/05/2024 Revised version: 17/09/2025
Accepted: 30/10/2025 Published: 25/02/2026

Abstract

In this paper we make the case for the further social anchoring of linguistic typology by illustrating recent methodological developments in the field of comparative research on language contact. We begin by discussing similarities and differences between sociolinguistics and language typology and focusing on the issue of the social anchoring of research on language variation and linguistic diversity. We argue that while sociolinguistics is socially anchored by definition, linguistic typology has so far abstracted languages from their social contexts due to the nature of macro-comparison and the poor availability of data on sociolinguistic environments. We suggest that greater social anchoring in large-scale comparative research on language structures is possible and can be achieved through integrating two general aspects of language use, relationality and context-dependency, into typological models of language variation and change. We illustrate how data collection on bilingual language ecologies embodies the notion of relationality and context-dependency.

Keywords: Sociolinguistics; language typology; language contact; language ecology.

1. Sociolinguistics, language typology and the social anchoring of language variation

Sociolinguistics and language typology are among the subfields of linguistics that share an interest in linguistic variation. Despite this shared interest, ever since they emerged as established domains of research in the 1960s (through the work of

William Labov for sociolinguistics and Joseph Greenberg for typology), the methodologies and research materials of these two fields have been quite different.[1]

A key idea in sociolinguistics is that language is a social institution and thus intertwined with different social phenomena. Much evidence has been accumulated showing how language use varies, for instance, across age and gender, how that variation may be harnessed for indexing social identity (e.g., Silverstein 2003), and how socially structured variation may lead to language change over time (e.g., Milroy & Milroy 1985; Croft 2001; Nevalainen & Raumolin-Brunberg 2017). Data is typically collected from individuals, analyzing their language production as well as some relevant social aspects hypothesized to structure production, such as age or gender and, more recently, community of practice (vid., Wenger 1998; King 2019). Macro-level analyses are relatively uncommon, although they appear to be somewhat more common in research on the sociology of language, which can be broadly defined as the study of the relationship between language and society with a focus on the societal aspects of linguistic behavior (Chen 1997).[2]

In language typology, an overarching aim is to better understand the world's linguistic diversity; that is, why the thousands of human languages spoken or signed across the globe are the way they are, what unites them, and how they differ from one another. Research in this field typically draws data from descriptive research materials, such as reference grammars. The focus is usually on different qualitative and quantitative aspects of linguistic patterns rather than on their behavioral profiles. Recently, however, corpus-based typological research that enables measuring usage frequency and assessing behavioral properties has been growing (e.g., Levshina 2019) thanks to the increased availability of multilingual annotated corpora, such as Universal Dependencies (Zeman et al. 2023), MultiCAST (Haig & Schnell 2023) and DoReCo (Seifart et al. 2024). Data from individual language users is rarely collected in typological research, although incipient examples do exist (e.g., Dingemanse et al. 2013).

[1] Francesca Di Garbo and Kaius Sinnemäki contributed to all aspects of the work, from the study design to the write-up. Eri Kashima contributed to the write-up of the work and the design of section 3.2.

[2] Sometimes sociology of language and sociolinguistics are treated as two separate fields despite their considerable overlap. Here we subsume sociology of language under sociolinguistics (following e.g., Bell 2013) but refer to the former where necessary for the sake of clarity.

These two research fields thus differ in various ways. One difference is that of scale. Sociolinguistics tends to focus on micro-level analyses that investigate variation at the level of individual members of a community (population of individuals). Language typology, on the contrary, tends to focus on macro-level analyses that study variation across languages (population of languages). In other words, while, in sociolinguistics, the units of comparison tend to be individuals and their sociological and linguistic profiles, in language typology, the units of comparison are languages or specific constructions in given languages, essentially viewed as shared communication systems.

Another major difference between these fields is in the social anchoring of research on language variation. What we mean by anchoring is whether research on language variation is informed by what is known about the social contexts in which variation is embedded; that is whether our models of variation and change do justice to the social nature of the linguistic phenomena observed.[3] Languages do not exist in a vacuum but in different sociocultural and geopolitical contexts that are in turn anchored in human behavior. Ideally, our models of language variation should account for those contexts to evaluate the extent to which human behavior, including linguistic behavior, may depend on those contexts. Because of its commitment to the individual language user and the social aspects of language use, social anchoring is a theoretical as well as a practical aim in sociolinguistic research. However, research in language typology tends to lack this kind of anchoring. Typological models are usually based only on language descriptions without a connection to the social contexts in which the described patterns occur.

[3] The notion of *anchoring*, as we use it in this paper, may somewhat overlap with that of *grounding*, as used in philosophy. In philosophy, the notion of grounding is understood as a metaphysical, noncausal relation between some aspect of reality that gives rise to or is the basis for some other aspect of reality: when A is grounded in B, A depends at least partially on B (see e.g., Correia & Schnieder 2012). Thus, saying that language is grounded in the social reality of speakers and signers is making a claim that some properties of language depend, at least to an extent, on the language users, albeit not causally but constitutively. While this may be a reasonable position when approaching some aspects of language, our interest in the relationship between language variation and the social context of language use concerns the probabilistic and possibly causal relationship between the two. We are specifically interested in whether language structures may adapt to the sociolinguistic context in which languages are learned and used (see Di Garbo et al. 2021; also Adamou 2021). We decided to use the more theoretically neuter term *anchoring* in order to avoid any potential pitfall related to using the more connotated term *grounding*.

Only in the past two decades have societal factors started to play a role in a research area now called sociolinguistic typology (e.g., Trudgill 2011; see the review in Sinnemäki, in revision). This research may integrate aspects of the language community into their models, such as size of the language community or intensity of language contact. But, otherwise, factors related to language users tend to play no direct role in data collection and analysis in language typology. Some such factors, such as cognitive processing preferences and learning constraints, may be used for explaining typological distributions but typically only in a post hoc fashion, rather than by being built into the framework (e.g., Lupyan & Dale 2010; but see research on efficiency, for instance Levshina 2022, where cognitive preferences may be used to at least generate hypotheses that are then tested with typological data). Furthermore, these factors seem to be limited mostly to individual cognition, such as memory limitations, rather than to social aspects of language use (see Kusters 2003 for an early example of this approach, focusing on second language learning and linguistic complexity in the verbal domain).

There are both historical and methodological reasons for this situation in language typology. As outlined by Bickel (2007), up until the late 20th century, typological research often focused on the question of what is possible in language. The aim was to find universal cognitive constraints, and typology was often seen as the flipside of research on universal grammar. Language was largely abstracted away from its social environment and attempts to correlate language structure and language-external factors were approached with strong reservations (see e.g., the discussion and references in Ladd et al. 2015: 227). Since roughly the 1990s there has been a steady movement into appreciating variation at all levels, giving typology a much more anthropological orientation (Bickel 2007). As a result, research in modern language typology tends to be probabilistic and embraces the fact that most typological features are geographically and genealogically distributed. Language universals are seen as structural pressures that affect how languages change over time rather than as cognitive constraints (e.g., Greenberg 1978; Sinnemäki 2010; Dunn et al. 2011; Bickel 2013). This “anthropological” shift with an appreciation of local variation has been a boon that has advanced research in the field (the journal *Linguistic Typology at the Crossroads*, with its programmatic effort to encourage debate around typology and its bordering disciplines, is a good example of this process). Along these lines many typologists also assume that the language system is not autonomous from language users but embedded in and affected by various language-external factors, such as

language use, social interaction and cognitive processing (e.g. Trudgill 2011; Hawkins 2014; Sinnemäki 2014; Pakendorf et al. 2021; Levshina 2022).

Despite this theoretical stance, linguistic patterns tend to be analyzed in typological research as if they were autonomous of language users. As mentioned above, language users enter the research design only post hoc, if they enter at all. Crudely said, and we are being self-critical here as well, especially in the field of large-scale quantitative typology, the data type that may be the most "social" seems to be the sheer geographical location of languages in terms of longitudes and latitudes, often represented as focal points. Geographical locations may then be used as measures of geographic distance or as a proxy for evaluating or controlling for the effects of language contact (see, for instance, Guzmán Naranjo & Becker 2022; Guzmán Naranjo & Jäger 2023). But language contact is founded on bilingual and multilingual interactions that are embedded in the social fabric of language use. It is also not clear how the geographical distance between language communities may be related to the social environment of language use more broadly, and this could in fact be turned into a research question on its own.

In this article, we build on the ongoing anthropological shift in language typology and argue that there are good theoretical grounds for taking this shift even deeper through a better social anchoring of research in the field. This social anchoring would involve, for instance, better addressing typical concerns in the sociology of language such as "who speaks what language to whom and when" (Fishman 1965), and thus producing descriptive data on the social environment of language use. In the next sections, we argue that such a shift in language typology is methodologically feasible and illustrate how it can be implemented concretely. We tackle the relationship between language structures and language use patterns by focusing on language contact dynamics and the impact of multilingual language ecologies on patterns of language structures. In language contact studies, there is a long tradition of research focusing on explaining the scenarios underlying the outcomes of language contact (see, for instance, Muysken 2013; Ross 2013). Thus, language contact phenomena provide a promising laboratory for investigating the social anchoring of linguistic diversity.

The paper is structured as follows. In Section 2, we discuss in more principled terms the conditions for increasing the social anchoring of typological research to base it more on naturalistic data. Two principles are introduced to this effect, what we call *relationality* and *context-dependency*. In Section 3, we demonstrate how we implement

these principles in a research design that is geared to capture contact-induced language change in bilingual language ecologies across the world. Section 4 ends the article with discussion and brief conclusions.

2. How to increase the social anchoring of research in language typology

2.1. On abstraction in language typology

In any field of research, comparison implies some level of abstraction. This is necessary in order to navigate individual manifestations of a given phenomenon while, at the same time, searching for patterns in the distribution of similarities and differences between data points (for an overview of the epistemology and history of the comparative method in the Humanities and Social Sciences, see Griffiths 2017).

In comparative linguistics, and language typology in particular, abstraction stems from at least two facts. Firstly, in the large majority of cases, the focus of comparison is not on languages as manifested through the repertoires of individual users, but rather on descriptive resources. These provide a reified representation of the functioning of said languages, what typologists refer to as *doculects* (Cysouw & Good 2013). Secondly, and related to the first point, given that the focus of comparison is on constructions as represented and described in a doculect, the generalizations stemming from large-scale typological research are inevitably limited to the linguistic properties of these constructions rather than encompassing the actual usage preferences of the population of individuals who associate themselves with a given language.

In sum, the population that is being compared in typological studies comprises a selection of doculects or, more specifically, a selection of constructions in a sample of doculects. Thus, even though, broadly speaking, many typologists conceptualize languages as communicative systems embedded in their own socio-historical, cultural and environmental ecologies, large-scale typological research tends to have a much broader and somewhat abstract focus, given that the object of study, languages and their constructions, are approximated through their documentary description, that is, doculects. Given this background, and compared to other fields of linguistic research, such as pragmatics, conversational analysis and sociolinguistics, the social agency of speakers probably cannot be operationalized in typology at the same level of detail. Typological data alone have thus far offered simplified representations of language

variation because they tend to leave out the social foundations of variation at the level of the individual.

However, this abstracting away from the social foundations of variation may also depend on the state of the art in data availability in language typology. Typology is about comparison, but comparison is possible only when sufficient and comparable data is available for doing so. Linguistic descriptions have been available for several hundred languages already for decades, making it possible to collect large typological datasets on linguistic features. However, descriptive and comparative work on the sociolinguistic environments of language use has been much slower and, thus far, it has also prevented language typology from greater social anchoring (Sinnemäki, in revision), However, the fact that, for instance, recent descriptive grammars sometimes provide actual data on the social correlates of structural variation at the community level, as in the work by Kluge (2017) on Papuan Malay (pmy; Austronesian, Nuclear Malayic), testifies of a growing interest in the social anchoring of language use.

Yet another sign of this ongoing paradigmatic shift towards a better integration between, on the one hand, language variation and change at the community level and, on the other hand, worldwide linguistic diversity is the work by Mansfield et al. (2023) on grammatical variation between closely related dialects in a typological perspective. They investigate how the social-signaling role of grammatical variation may contribute to linguistic divergence and diversification. This is done using a sample of 42 languages across the world and based on documented descriptions of dialectal variation in these languages. The study finds that dialect differentiation between closely related varieties in close contact is mostly carried out by what they call *form-variables*. This is when dialectal difference is manifested by variant forms of the same morpheme across dialects, such as the negative auxiliary/copula variation in spoken varieties of English[4] (*It **isn't** right* vs. *it **ain't** right*). This variation is interpreted as a sign of the strong social indexing role that linguistic divergence plays in situations of dialect contact. The method developed by Mansfield et al. (2023) represents a groundbreaking innovation as it shows that dialectal differences, which are usually investigated in variationist sociolinguistic and dialectological studies of individual speech communities, can be also studied from a large-scale typological perspective.

In this paper, we argue that increasing the social anchoring of large-scale typological investigations is also possible when comparing genealogically unrelated languages. However, we argue that this step requires some further level of

[4] eng; Indo-European, Germanic.

abstraction. We therefore propose a new type of heuristics in the study of linguistic diversity and one which is anchored to the relational and context-dependent nature of linguistic phenomena, as presented in the next subsection.

2.2 The relational and context-dependent aspects of language use

We propose two general principles related to the social aspects of language use and argue that they both provide a viable starting point for increasing social anchoring in language typology. We also demonstrate how these principles could be built into a typological approach in practice. These two principles are (1) the relational and (2) the context-dependent nature of language use. First, humans have a disposition to interact and cooperate with one another, to share intentions, and to form coordinated group activities (e.g., Tomasello 2010; among many others). All these aspects are closely tied to language and its use: one of the main functions of language is communication, which by definition is relational, interactional and cooperative. Second, language use is strongly context-dependent, which means that human communicative practices vary depending on the situation, on the audience, and even on eavesdroppers. This context-dependency is the basis for much linguistic variation, such as different styles, genres and registers (e.g., Biber 1988; among many others).

In our typological research, we have turned these two general social aspects of language use into two interrelated research foci about language ecology, as in (1). The term language ecology here broadly refers to the interaction between language and its environment (Haugen 1972), but we limit our discussion to the social and sociohistorical aspects of the environment.

(1) Our operationalization of language ecology
a. Relationality → Focus on bilingualism and language contact
b. Context-dependency → Focus on variation in language ecology and in language use

In terms of the relational aspect of language use, we focus on bilingualism and language contact, that is, how language changes when interaction takes place between people speaking different languages. We approach the context-dependent aspects of language contact in the following two ways, both focusing on variation. First, we research the social aspects of the bilingual language ecology in different

social domains, such as the family or the occupational context, and not simply overall in the community (e.g., Fishman 1965; Di Garbo et al. 2021). Second, we attempt to capture linguistic diversity in the spirit of multivariate typology, which proposes fine-grained typological variable design as a way of integrating information about language-internal variation in large-scale comparative research on the world's languages (e.g., Witzlack-Makarevich et al. 2022). Through this demonstration, we aim at illustrating what contribution this newly developed approach brings to the typologist's toolkit, by making a step forward towards a principled understanding of linguistic structures as embedded in their language ecologies.

Applying these two principles concretely would mean creating datasets that incorporate as much information as possible about language-internal variability in the distribution of linguistic features. It also means factoring in characterizations of the sociolinguistic environments of language communities. Capturing language-internal variability is a way of getting at the relational aspect of language use, that is to the fact that linguistic variants are constantly negotiated through interactions between individuals and the representations that people build of their interlocutors (relationality). In our work, we specifically focus on developing and testing methods that can allow us to detect how bilingual language use affects the emergence and development of linguistic variants. For instance, if several alternative strategies of encoding are attested for one and the same linguistic feature, could any of these have emerged as a result of contact with neighboring communities? At the same time, characterizing the social embeddings of linguistic interactions provides a way of assessing how the distribution of linguistic variants may also depend on the specific socio-cultural contexts in which these interactions are situated (context-dependency). For instance, can language dominance or language ideologies tell us anything about the type of linguistic variants that are more likely to emerge in contact situations?

The relational and context-dependent principles of language use have been the objects of recent discussions in small-scale multilingualism research (for an overview see Pakendorf et al. 2021). Linguistic identities in small-scale multilingual societies are characterized as multifaceted and sensitive to contexts of interaction. These contexts of interaction are highly localized and can be based on factors such as place (e.g. for speaking the languages associated with specific places see Merlan 1981 for Australia and Döhler 2018 for southern New Guinea) or relationship with interactant (e.g. for speaking to in-laws, see Fleming 2011 and for speaking to clan members, see

Garde 2008; Suokhrie 2016; for speaking a village language of communication, see Gumperz & Wilson 1971).

Additionally in those small-scale multilingual societies where the linguistic repertoire of language users coincides with a pool of closely related varieties, it has been observed that minimal structural differences across varieties may often convey strong social meanings, associated with one or the other speaker community. Lüpke (2022) illustrates these processes of socially charged linguistic divergence in the nominal classification systems of the languages of Lower Casamance (Senegal). These closely related varieties typically display the same inventories of noun class distinctions from a formal point of view. However, the assignment of individual nouns to such classes differs across varieties and these differences actually index people's 'belonging' or 'identifying' themselves as members of one or the other community. Examples are shown in Table 1.

Baïnounk Gubëeher	**Joola Kujireray**	**Meaning**
bu-óóg/i-óóg	*fu-bah/ku-bah*	'baobab fruit(s)'
bu-gof/i-gof	*fu-how/ku-how*	'head(s)'
bu-koor /-i-koor	*e-suh/si-suh*	'village(s)'
bu-deen	*e-baŋ*	'putting'

Table 1: Mismatches in noun class assignment across two closely related varieties of Lower Casamance (Senegal). Examples taken from Lüpke (2022)

In Baïnounk Gubhëeher (gube1234; Altlantic-Congo, North-Central Atlantic)[5] and Joola Kujireray (bkj; Altlantic-Congo, North-Central Atlantic), two languages spoken in Lower Casamance, the nouns class markers *bu-/i-* (Baïnounk Gubëeher) and *fu-/ku-* (Joola Kujireray), respectively, are semantically equivalent cognate classes that are used for the classification of round things. As suggested by the examples in Table 1, the same noun class markers are used in these languages for some words, such as the nouns for 'baobab fruit(s)' and 'head(s)'. However, nouns that are assigned to this semantically motivated class in one of these varieties may end up being assigned to other classes in the other, such as the nouns for 'village', which belongs to class *bu-*

[5] This language does not have an ISO-code; the Glottocode is used instead.

/i- in Baïnounk Gubëeher but is assigned to class *e-/si-* in Joola Kujireray. Such differences in class assignment across closely related varieties may then be used for indexing people's 'belonging' to one or the other community. For analogous examples, in a different small-scale multilingual setting within Western Africa, the Cameroonian Grassfields, see also the recent study by Di Carlo & Good (2023).

Our claim is that the relational and context-dependent aspects of language use, which have already been identified as relevant for understanding and modelling types of multilingualism, are widely applicable to any situation of contact between users of different languages. Accounting for these aspects of language use in typological studies could thus provide a starting point for increasing the social anchoring of the crosslinguistic generalizations that are made in these studies.

Once this possibility is acknowledged, the onus is to make explicit and workable steps towards turning these observations into implementable methodologies. In particular, in order to turn this fluid and versatile representation of the workings of language into something that can be compared across time and space, some form of reification becomes necessary. For instance, while in linguistic research there is a growing call to shifting the focus from *languages* as compartmentalized entities to *languaging* as the constant negotiation of communicative repertoires in context (see, e.g., Lüpke 2024), comparing linguistic structures, as used by a population, still requires some sort of schematization of the phenomena being compared, be it a set of specific constructions or a holistic representation of a linguistic code. Importantly, psycholinguistic studies suggest that reified representations of languages and linguistic practices also have some kind of psychological reality (see, for instance, Berthele 2021). That is, a linguistic code may be a clearly identifiable object for multilingual speakers, and the separability of the code(s) may be what enables languaging in the first place.

In section 3 we illustrate how the relational and context-dependent nature of language use may be operationalized typologically in the crosslinguistic study of language contact.

3. Language contact in its bilingual language ecology: some illustrations

In the previous sections we argued for the theoretical importance of increasing social anchoring in language typology. We discussed the methodological requirements for doing so and the reasons related to the lack of such anchoring in earlier typological research. We suggested that the relational and context-dependent nature of language

could be important starting points to address when linking language typology to the social aspects of language ecology.

Here we will illustrate how these principles have been built into a new typological research design stemming from research conducted during the ERC Starting Grant project GramAdapt (e.g., Di Garbo et al. 2021). While this illustration relies on research conducted or published in the context of this larger project, the discussion of the principles of relationality and context-dependency is original to the present paper. We discuss how each of the two principles has been built into the broader research design of the project, and how ensuing methodological issues have been addressed.[6]

3.1. Relationality and context-dependency in GramAdapt linguistic data

To research the relational nature of language use, a suitable linguistic phenomenon is needed that could be feasibly compared across languages. We argue that language contact and bilingualism offer one such area. Contact and bilingualism are relational from the outset, because in bilingual ecologies two or several populations of individuals speaking different languages may interact with one another and may thus also influence one another's linguistic behavior.

To assess such influences across languages, we have developed a new typological approach to language sampling, which is geared to make inferences about contact-induced change (see Di Garbo & Napoleão de Souza 2023 for details). Languages are selected in pairs based on prior evidence of interaction between the language communities of interest. The primary language of interest in any given pair is identified as the Focus Language, and its contact language is the Neighbor language. The pairs of Focus and Neighbor Languages form a "test case" that lets one zoom in on the relational nature of contact. An example of a test case is the contact between Alorese (aol; Austronesian, Bima-Lembata) and Adang (adn; Timor-Alor-Pantar, Nuclear Alor-Pantar) in the East Nusantara region of Eastern Indonesia (see Figure 1). Notice that Alorese is in contact with other Alor-Pantar languages spoken in the Pantar Islands and its islets. Here we follow the work by Moro (2021) who focuses on

[6] There are also other approaches to implementing the relational nature of language use into typological analysis, for instance, in pragmatic typology (see e.g., Floyd et al. 2020 and Rossi et al. 2020). However, those approaches focus on interactional dynamics and thus assume finer-grained data than what is possible in the context of the approach presented here, which is geared towards studying structural features of languages.

the Alorese-Adang contact scenario because of its neater dynamics in comparison to the other neighboring languages. We refer to Moro's paper for an in-depth study of the multilingual patterns of the Alorese community.

To analyze whether contact with the Neighbor language has led to changes in the Focus language, a "control case" is also selected; a language that is closely related to the Focus language but that has not been in contact with either it or the Neighbor language (Di Garbo & Napoleão de Souza 2023). As an example, the western varieties of the Austronesian language Lamaholot (slp; Austronesian, Bima-Lembata) would serve as a reasonable control case, since it is known that Alorese split off from western Lamaholot no later than roughly 600 years ago (Klamer 2012). At best, such a control case, or Benchmark, may approximate the state in which the Focus language was prior to its contact with the Neighbour (e.g., Sinnemäki & Ahola 2023; Sinnemäki et al. 2024). In this paper, we choose the Leiwong dialect of Lamahalot, described by Nishiyama & Kelen (2007), as a Benchmark.

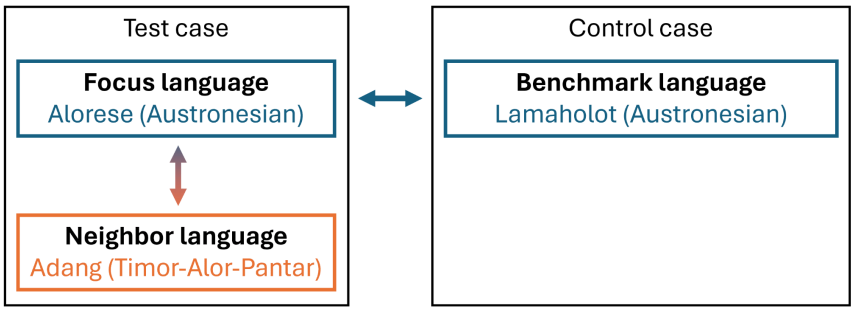

Figure 1: "Test case" and the "control case" according to the sampling scheme of Di Garbo & Napoleão de Souza 2023).

This approach to language sampling shifts the attention from viewing languages in isolation to incorporating the relational nature of language contact into the model. The triplet Focus-Neighbor-Benchmark also enables us to draw dynamic inferences about the outcomes of language contact (for earlier studies in dynamic typology see, e.g., Maslova 2000; Maslova & Nikitina 2007; Cysouw 2011; Bickel 2013). Since more than one language is selected from the same family, it is possible to draw inferences about type shifts, that is, how the Focus language may have changed in the contact situation, and as compared to the Benchmark sister language.

Focusing on language contact from a worldwide comparative perspective may raise some methodological issues. For instance, a basic requirement in quantitative

research is that datapoints are independent of one another (see Winter & Grice 2021 for a recent discussion). Yet, when languages are sampled in contact pairs, they cannot be considered fully independent of one another. The Focus and the Benchmark also come from the same language family and therefore are not independent of one another. This point is broadly shared by all diachronically minded approaches to typological research, such as dynamic typology (e.g., Bickel 2013): when typological universals are interpreted as diachronic pressures to type shift, in order to assess the dynamics of these type shifts, it becomes necessary to select two or more languages from the same family. Classical statistical tests therefore cannot be applied to the data sampled in such a way.

In our approach, the datapoints are not linguistic features of individual languages but potential contact-induced type shifts (and "non-shifts", see below) in the Focus Language which are identified and analyzed by taking into account what is attested in the respective Focus-Neighbor-Benchmark triplet. Our approach is thus inherently dynamic in that it understands language structures as constantly evolving. In order to capture these patterns of evolution, the linguistic variables of interest are situated in a broader context by taking into account what they could have become if contact-induced change occurred (i.e. given structure as realized in the Neighbor language), and by comparing them with analogous structures as attested in a close relative (the Benchmark).

As an illustration, let us consider the order of the possessor and possessum in alienable possessive constructions in our example set (Alorese-Adang-Lamaholot). In the Benchmark Language Lamaholot, the possessor follows the possessum, as in (1). In the Neighbour Language Adang, the possessor precedes the possessum, as in (2). The Focus Language Alorese has developed the possessor-possessum order, as in (3), possibly via contact with Adang. It is thus the type shift from the possessum-possessor order to the possessor-possessum order in Alorese that is the independent data point to count in these constructions.

(1) Benchmark: Lamaholot (Bima-Lembata, Austronesian; Nishiyama & Kelen 2007: 23)
lango *go'en*
house 1SG.POSS
'my house'

(2) Neighbour: Adang (Timor-Alor-Pantar; Haan 2001: 147)
n-ɔ *baŋ*
1SG-GEN house
'my house'

(3) Focus: Alorese (Bima-Lembata, Austronesian; Sulistyono 2022: 106)
go *uma*
1SG house
'my house'

A non-shift, in turn, is simply a situation when the Focus Language has retained the allegedly inherited feature. The feature is also attested in the Benchmark Language but not present in the Neighbor language. For instance, the nominative/accusative form of the first-person pronoun in the Benchmark Lamaholot is *go*, while its genitive form is *go'en* (Nishiyama & Kelen 2007: 13). The first-person genitive pronoun in the Neighbour Language Adang is *nɔ* (or *ne*; Haan 2001: 149). In the Focus Language Alorese, the first-person form of the pronoun is *go* and this form is used for subject, object and possessive functions. In other words, Alorese seems to have retained the form of the first-person singular pronoun *go*, although it has also lost its genitive form.

As long as the triplets are independent of one another, it makes sense to assume that the distribution of type preferences gleaned from the triplets are also independent of one another. Thus, drawing inferences from the triplets does not preclude the use of statistical methods. Even if one was unwilling to use classical statistical methods for such data, logical inferences can still be made about such distributions with the help of Monte Carlo methods (e.g., Janssen et al. 2006).

Integrating the relationality principle to sampling has thus significant repercussions to data collection: to make inferences about language contact, linguistic data need to be collected from each contact pair but also from each Benchmark Language. This data collection can be done through standard practices in language typology; for instance, by analyzing descriptive linguistic data collected from reference grammars.

In addition to factoring in the relational aspect of language use, by looking at pairs of languages in contact, our method also lets us explore contextual variability in linguistic structures. We achieve this by working with a coding design that is purposely geared towards exploring patterns of language internal variation for any

given linguistic variable (Sinnemäki et al. 2024). This coding design is inspired by the principles of multivariate typology (Witzlack-Makarevich et al. 2022), whereby linguistic structures are explored through fine-grained questions that contribute to depicting and characterizing variation. For instance, with respect to nominal number, we do not just look at whether a given language marks the plural, but whether the plural is marked on nouns, pronouns of various types, adnominal modifiers, verbs etc. And, with respect to nominal number marking, we do not just ask whether this is suffixal or prefixal but whether different types of markers (suffixes, prefixes, stem alternations, reduplication) may occur on nouns or through agreement (Di Garbo & Kapellis 2025). Through this procedure, we can detect whether languages exhibit different strategies in different domains and whether this speaks of any ongoing change. We indeed find internal variation in the distribution of linguistic features of the Focus languages. When comparing this variation with the constructions attested in the respective Neighbor and Benchmark languages, these patterns of internal variation can, in some cases, be interpreted in terms of contact-induced type shifts.

Coming back to the Alorese-Adang contact pair as an illustration, similarly to its Neighbour Adang, Alorese marks nominal plurality by means of a plural word, *hire*, which is historically derived from third person plural pronouns. This is shown in (4) and (5). Conversely, the Benchmark Lamaholot does not have plural words, but nominal plurality is optionally marked through reduplication, as shown in (6).

(4) The plural word *hire* in Alorese (Moro 2018: 184)

məsia	*hire*	*ke*
person	PL	DEM.PROX

‘these/the persons’

(5) The plural word *nun* in Adang (Haan 2001: 122)

pen	*ti*	*mat*	*nun*	*ʔa-bɔʔɔi*
Pen	tree	big	some/several	3.OBV-cut

‘Pen cut some big trees.’

(6) Plural reduplication in (Lewoingu) Lamaholot (Nishiyama & Kelen 2007: 210)

inamvlake-inamvlake	*svga-ka*	*urin*
man-man	came-3PL	late

‘Men came late.’

A few reduplicated nominal stems also exist in Alorese, but their non-reduplicated bases are no longer attested. This suggests that Alorese also used to mark plurality through reduplication as does its sister language Lamaholot, but this strategy was later replaced by the emergence of plural words, as in the Neighbor language Adang. In this case the inferences drawn from our coding method can be backed up by existing literature. Moro (2018) demonstrates that the grammaticalization of the Alorese plural word is indeed the result of a contact-induced type shift in the wider historical context of contact between Austronesian and Alor-Pantar languages. Additional examples of language internal variation in nominal number systems which are suggestive of ongoing contact-induced type shifts in the languages of our sample can be found in Di Garbo & Kapellis (2025).

These examples illustrate how we address contextual variability in linguistic structures through fine-grained analyses of the structures attested in the Focus languages and then comparing them with data from the respective Neighbor and Benchmark languages. We argue that capturing language-internal variation in this way is one of two possible ways of exploring context-dependence in contact-induced variation from a typological perspective. Another way, which we have not yet implemented in our own work, is to compare language structures in closely related dialects with different contact profiles to test whether dialectal variation can be eventually explained as a function of proximity or degree of contact with a genealogically unrelated neighbor (see Mansfield et al. 2023 for a recent advance in dialect typology, also discussed in section 2.1) These two approaches make the typological study of language variation and change in contact situations more dynamic; that is, by capturing language-internal variation through comparisons of genealogically unrelated contact pairs on the one hand, and by comparing dialect pairs with different contact profiles on the other. As we hope to have shown, the two approaches have the potential to significantly contribute to increasing the ecological validity of typological generalizations.

3.2 Relationality and Context-dependence in GramAdapt sociolinguistic data

In our approach to contact and bilingualism, we also factor in the sociolinguistic aspects of the bilingual language ecology. Collecting such data is, however, hindered by poor data availability. There is little descriptive data available on sociolinguistic environments in general, and on bilingual language ecologies in particular.

Descriptive grammars often contain sections on language ecology, but their extent varies and their usefulness for evaluating bilingual language ecologies in particular may be low. Other initiatives, such as the articles in the Language Context section of the journal *Language Documentation and Description* may contain more useful data but tend to focus on a community's use of a single language. Overall, the descriptive status of sociolinguistic environments of languages is rather low.

For this reason, we developed a sociolinguistic questionnaire within the project, with the aim of eliciting fine-grained descriptions of the Focus-Neighbor contact profile at a particular point in time (see Kashima et al. 2025 for an overview of the questionnaire design). The point in time was defined as when there were the most opportunities for interaction between the Focus and Neighbor language speakers. The sociolinguistic aspects of contact are thus analyzed between selected individual languages within the "test case". The Benchmark Language was selected specifically because it has not been in contact with the Focus or the Neighbor and hence no social contact data was collected on it. The questionnaire was directed to specialists such as documentary linguists and anthropologists who are well familiar with the Focus-Neighbor contact situations. Here, we focus on the methodological choices that we made in order to tie this questionnaire to the relational and context-dependent aspects of language use and also illustrate initial results. For the purpose of this illustration we draw data from Kashima et al. (2023), which is a dataset of 34 responses to the sociolinguistic questionnaire, featuring 34 contact scenarios from around the world and partially overlapping with the sample presented by Di Garbo & Napoleão de Souza (2023). Figure 2 shows the distribution of the contact scenarios.

Contact research was at the heart of the first attempts to increase the social anchoring of typology when sociolinguistics and language typology were first being integrated more than 20 years ago (see Sinnemäki, in revision). The most commonly incorporated sociolinguistic data in these studies were population size and some broad qualitative assessment of the intensity of contact (e.g. Sinnemäki 2009; Lupyan & Dale 2010; Bentz & Winter 2013; Sinnemäki & Di Garbo 2018). Thus, although some aspects of social contact were integrated in typological models, these factors were analyzed on a general level and without investigating contact relations between specific languages, that is, in an essentially non-relational way (see below). For instance, Bentz & Winter (2013) collected data on the proportion of non-native

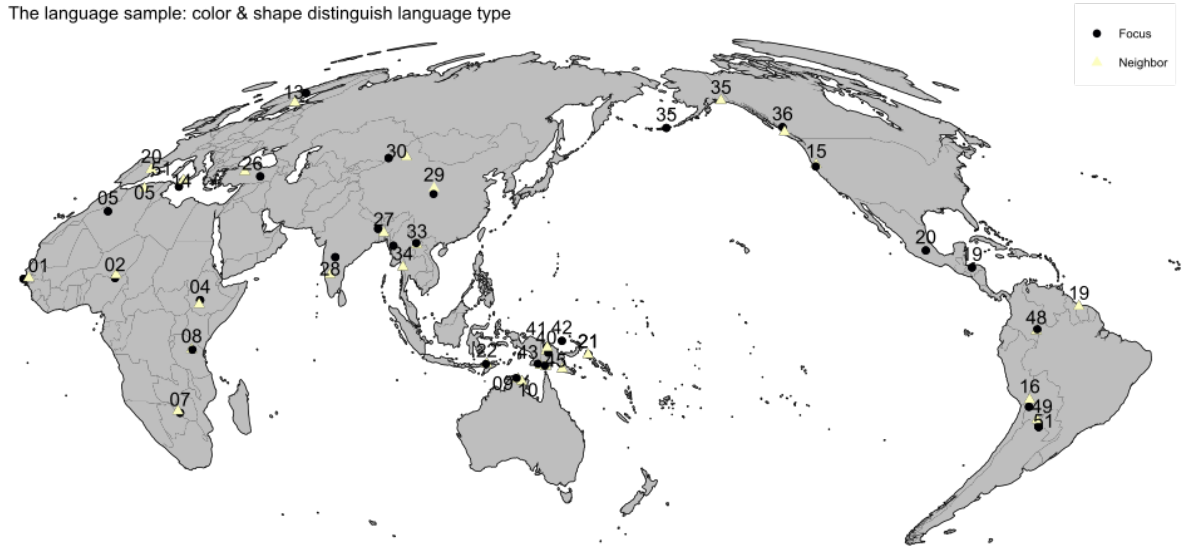

Figure 2: the 34 contact scenarios of the sociolinguistic dataset. Focus languages are represented as black circles and Neighbor languages as beige. Contact pair IDs are also plotted.

speakers in whole speech communities (see also Trudgill 2011). Such general-level demographic data provide information about the overall extent of bilingualism in the community and could potentially be connected to mechanisms of language change that depend on such general-level phenomena. If, for instance, the speaker population is accustomed to many people from different backgrounds learning their language, this could mean that the native population might also be accustomed to producing foreigner-directed speech, which is one of the mechanisms of language change assumed to lead to simplifications (e.g., Kusters 2003; Trudgill 2011; Bentz & Winter 2013; Berdicevskis 2020).

A concrete example of a non-relational way of describing a bilingual ecology could be the phrasing given in (7a) while (7b) illustrates how this example could be phrased in terms of the specific contact situation between Alorese and Adang.

(7) Non-relational ways of describing language contact
a. “Speakers/signers of language X are typically bilingual.”
b. “Speakers of Alorese are typically bilingual.”

This statement could then prompt a binary yes/no response, a set of Likert-scale type predefined values ranging from "very uncommonly" to "very commonly", or perhaps even a numerical value that would represent the extent of bilingualism in the community. Note that such a non-relational variable abstracts away from the specific contact relations between language communities and may result in loss of relevant information about the contact ecology.

In our approach, we build the relational nature of language contact into the model in terms of the sociolinguistic aspects of contact. We thus assume that contact-induced changes may be affected by the particular Focus-Neighbor contact ecology as well as by the properties of the Neighbor language. This emphasis on the relational aspect of the bilingual ecology implies developing ways to tackle the relational nature of a given contact situation, which is often quite straightforward. For instance, a non-relational variable, such as (7a), can be easily turned into a relational one by zooming in on a particular contact scenario, as in (8a). Examples (8b) and (8c) illustrate how these variables could be phrased in terms of our example test case involving Alorese and Adang.

(8) Implementing the relational aspect of language contact
a. "Speakers/signers of language X are typically bilingual in language Y."
b. "Speakers of Alorese are typically bilingual in Adang."
c. "Speakers of Adang are typically bilingual in Alorese."

If both linguistic and sociolinguistic data were collected on language X and Y, this would enable making inferences about how the social contact between the two language communities has potentially affected the languages in that contact scenario (see more in Sinnemäki & Kashima, forthcoming).

We illustrate how the relational aspect of social contact was built into our questionnaire, by drawing one example from Kashima et al. (2023). This example question asks about language attitudes towards linguistic transfer. A non-relational variable probing language attitudes to linguistic transfer could ask what the Focus language speakers' attitude to lexical or grammatical borrowing are on a general level, regardless of the source contact language. However, such attitudes often depend on the language at stake, and, for this reason, it makes sense to ask about attitudes in a relational way, that is, by focusing on specific contact pairs. For example, the notable difference in the prevalence of Swedish (swe; Indo-European, Germanic) vs

Russian (rus; Indo-European, Slavic) origin loanwords in Finnish (fin; Uralic, Finnic) throughout the twentieth century (Cronhamn 2018) is in part explained by the more positive attitude that Finns historically had towards Sweden and Swedish rather than towards Russia and Russian. Question ID OI6 in Kashima et al. (2023) thus asks the question in (9).

(9) Example question from the sociolinguistic questionnaire
"What are the Focus language speakers' attitudes towards linguistic transfers from the Neighbor language, such as lexical or grammatical borrowing?"

The predefined responses to this question are on a Likert scale and range from "Very negative" to "Very positive". The responses for 34 Focus-Neighbor pairs are summarized in Figure 2.

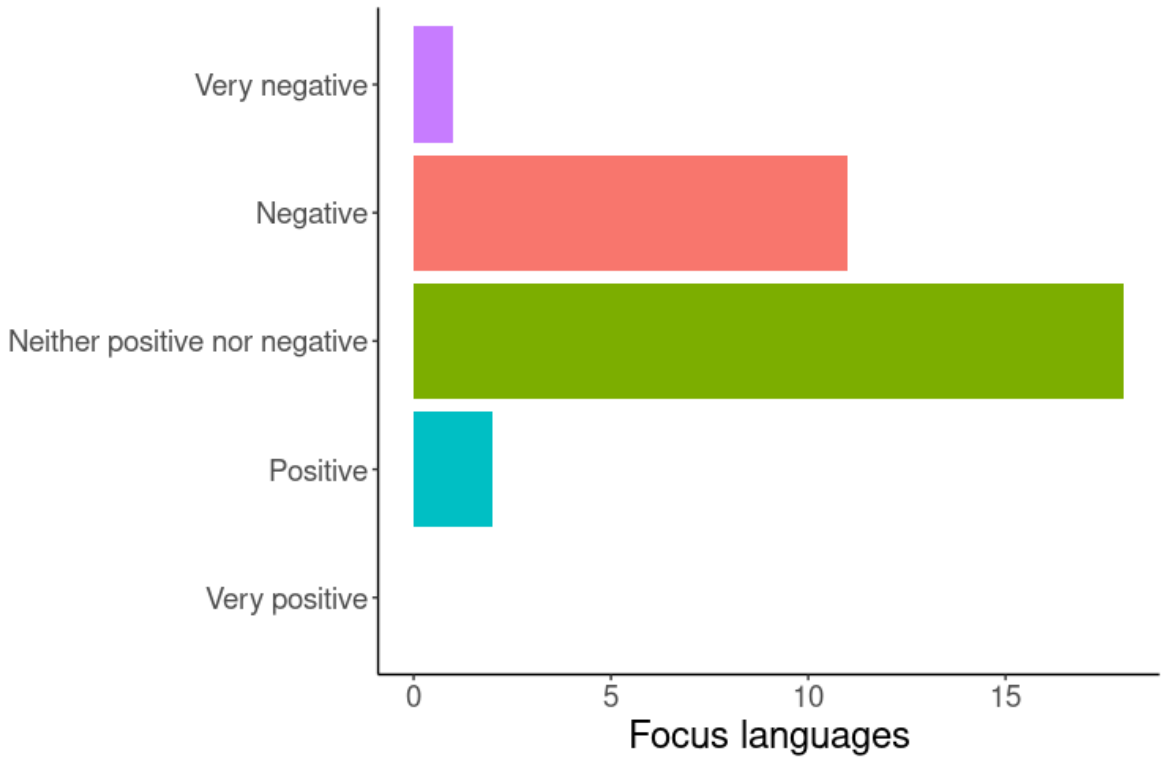

Figure 2: Distribution of responses about attitudes to linguistic transfer from the Neighbour Language to the Focus Language (data from Kashima et al. 2023).

Based on the specialists' responses, the Focus language speakers quite typically have either an indifferent or somewhat negative attitude to linguistic transfer from the Neighbor language. In other words, based on elicited specialists' assessments, it

appears that Focus language speakers would rarely consider lexical or grammatical borrowings from the Neighbor language as a good thing, although this response is not altogether absent in the data either.

As for the context-dependent factors of social contact, we asked questions about the social context of the contact across six social domains. The six social domains used are shown in (10) (see Kashima et al. 2025 for further details).

(10) List and definition of the GramAdapt social domains:

- Trade: concerning transaction of goods;
- Family and Kin: concerning relationships within the household;
- Local Community: concerning relationships in the private sphere beyond the household;
- Social Exchange: concerning relationships of non-transactional exchange, for example practices of ceremonial exchange;
- Knowledge: concerning relationships in the sphere of formalized learning (schooling and religion being the prototypes);
- Labor: concerning relationships in the sphere of production.

In most cases the same questions were repeated across all six social domains. This helped us to get an understanding of how the social aspects of contact between the Focus language speakers and Neighbor language speakers varied across social contexts.

We illustrate this point by taking another example from Kashima et al. (2023) that shows how the context-dependent aspect of social contact was built into a question asking about the occurrence of contact between the Focus and Neighbor group people. Instead of asking whether there was contact between the Focus and Neighbor people to begin with, we asked whether there was contact between the Focus and Neighbor people in each of the social domains. (11a), adapted to our example set Alorese-Adang in (11b), illustrates what this question looks like.

(11) Implementing the context-dependent aspects of social contact

a. Is there social contact between the Focus and Neighbor people in the domain of trade/labor/

b. Is there social contact between Alorese and Adang speakers in the domain of trade/labor/…

As is clear from this example, we often elicited the context-dependent and relational aspects of contact through one and the same question; context-dependency stems from anchoring the question in the individual social domains whereas relationality is related to the fact that contact is always framed from the perspective of the Focus-Neighbor interaction.

Figure 3 summarizes the occurrence of contact across the six social domains for 34 Focus-Neighbor pairs for which we have data on (Kashima et al. 2023). Based on the distribution of responses in Figure 3, there seems to be some evidence for a hierarchy of bilingual interaction across social domains. The hierarchy is summarized in (12).

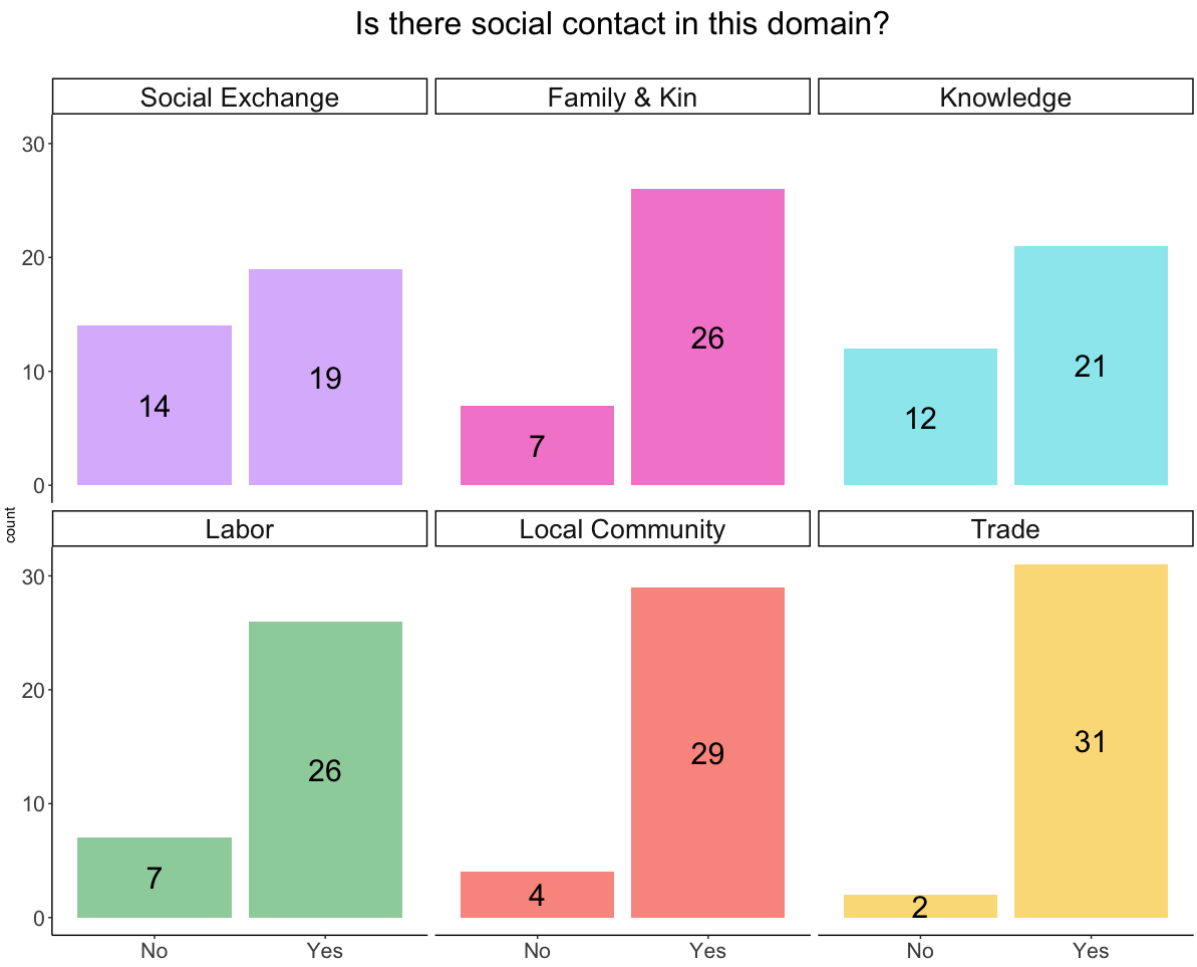

Figure 3: Distribution of responses on social contact between the Focus and Neighbour across social domains (data from Kashima et al. 2023).

(12) TRADE ≥ LOCAL COMMUNITY ≥ FAMILY & KIN ≥ LABOUR KNOWLEDGE ≥ SOCIAL EXCHANGE
97% 94% 82% 91% 82%

What this hierarchy states is that if there is contact in the domain of social exchange, there is likely contact in the other domains up in the hierarchy as well. However, if there is contact in the domain of trade, that does not yet tell whether there is contact in any other domains down the hierarchy. In our dataset, contact is thus most likely to occur in the domain of trade, which may also suggest that this is the context where contact relations between different language communities first develop. Conversely, language contact in the domains of knowledge and social exchange is the least typical and seems to first require contact in at least some other social domains higher in the hierarchy. The figures below the hierarchy reflect the percentage of the sampled sets in which the predictions of the hierarchy hold. For instance, only in six sets out of 34 are Focus and the Neighbor communities in contact in the domain of social exchange without being in contact in the knowledge domain. In other words, the prediction of the hierarchy for those two domains, i.e. contact in social exchange only if contact in knowledge exchange, holds in 27 out of 33 sets (≈ 82%). Overall, the predictions of the hierarchy hold to a very high degree (from 82% to 97%).

The sociolinguistic dataset stemming from our questionnaire responses contains data on roughly 200 variables on each set, totaling roughly 12,000 datapoints. The dataset makes it possible to answer various questions related to language contact and its social foundations, including questions about a particular contact scenario but also across contact scenarios. The list in (13) is a selection of possible general questions that can be potentially asked through this questionnaire.

(13) A sample of questions enabled by the GramAdapt Social Contact dataset

- Which sociolinguistic factors contribute the most to contact-induced change (cf. Thomason & Kaufmann 1988)?
- Does the rate of language change in contact situations depend on the social domain(s) in which contact takes place (cf. Greenhill et al. 2018)?
- Does the importance of language to group identity lead to greater divergence between languages (Braunmüller et al. 2014)?
- Does population size have any effect on contact-induced change and if yes, is it indirect so that it approximates the effect of some other social factors (Sinnemäki & Di Garbo 2018)?
- Is it possible to predict which sociolinguistic factors lead to simplifications or complexifications in contact situations (cf. Trudgill 2011)?

- Which social aspects of contact together affect contact-induced change (Trudgill 2011)?
- Does the extent of contact effects vary depending on whether children participate in contact dynamics (Trudgill 2011)?
- Are contact-induced changes in morphology driven more by changes in syntax than by aspects of the bilingual language ecology (Sinnemäki 2020)?

In the case of the Alorese-Adang contact pair which we exemplified in the previous section, sociolinguistic and historical data collected through the questionnaire (Moro & Sulistyono forthcoming) suggest that two main types of contact dynamics might have contributed to shape the patterns of contact-induced change attested in Alorese under the influence of Adang. Under the first scenario, possibly ongoing until the Dutch colonial period, the Adang dominated the Alorese. The Alorese-speaking population was bilingual in Adang and this situation of symmetrical bilingualism likely led to the patterns of grammatical restructuring discussed in 3.1, such as the grammaticalization of the plural word *hire* or the restructuring of adnominal possession patterns. After the start of the Dutch period, the Alorese prestige increased, which contributed to reshaping the extent of bilingualism from symmetric child-based bilingualism to Adang's adult-based bilingualism in Alorese. Language change processes stemming from this later type of contact settings mostly affect Alorese verbal morphology, which is undergoing considerable simplification under the influence of Adang L2 speakers (see Moro & Sulistyono forthcoming and references therein). This example clearly illustrates how the sociolinguistic data collected through the questionnaire may help elucidating the sociohistorical correlates of linguistic change and how these unfold through the history of a speech community.

4. Concluding remarks

In this paper, we tackled the broad question of how to define and account for "naturally occurring" data in linguistic typology by discussing the social anchoring of large-scale comparative research on language structures. We proposed that focusing on the relational and context-dependent properties of language is one way to investigate the social correlates of the distribution of linguistic diversity, thus ultimately boosting the social anchoring of typological generalizations.

We illustrated this point by presenting the research methodology of the GramAdapt project which has recently proposed a new way of approaching the dynamics of language change and linguistic diversification of the world's languages, focusing on language contact as a case in point. This novel approach intervenes on all key aspects of crosslinguistic research, from language sampling to data coding and statistical testing.

As shown in early sections of this paper, the GramAdapt typological and sociolinguistic datasets consist of observational data based on expert judgments, as is still typical of large-scale comparative research on linguistic diversity. Yet, the toolkit developed by this project strives to account for the relational and context-dependent aspects of language use at all stages of data collection and analysis. The sampling unit is eminently relational as it is constructed around documented contact scenarios between Focus and Neighbor languages, on the one hand, and proven genealogical relations between Focus and Benchmark languages, on the other. In addition, linguistic and sociolinguistic data are collected with the aim of capturing context-dependencies, both in terms of language-internal variation in the occurrence of individual linguistic features, and in terms of sociolinguistic variation with respect to observed dynamics of language use, language attitudes and ideologies across social domains.

Our approach to language variation and change is that of dynamic typology, whereby language universals are understood as universal pressure for type change. The time window for testing transition probabilities between linguistic types is the one provided by the comparative method. If we see a bias or a preference within that window, we can then extrapolate outside that window, other things being equal (Bickel 2013). But how justified is this extrapolation? In other words, have the factors that pressure languages to change remained the same even within this time window?

Languages are spoken in certain sociohistorical contexts, and many of those contexts have changed dramatically over the past few millennia (invention of agriculture and sedentary lifestyle, industrialization, urbanization, mass literacy, population movements, etc.). Research in (post-)colonial contexts squarely point to the coercive powers of colonial state infrastructure as responsible for societal upheavals (e.g. Givón 1971; vid. Yakpo 2020) which consequently act as pressures leading to language shift, death, and possibly the emergence of creoles and mixed languages. The linguistic consequences brought about by (forced) use of a state-sanctioned language, (forced) literacy and (forced) movement of people are well documented across the globe. We know less, however, about linguistic changes that

occur in non-colonial contact situations, and we are certainly still in the early days of developing a global-historical understanding of what types of pressures beget what kinds of linguistic changes – if there is indeed such a clear relationship. Our method partially allows us to tackle such future endeavors with the limits of its design principles.

Finally, the goal of the methodological shift that our methods propose and instantiate is not only that of getting new, groundbreaking answers to questions about the nature of human linguistic behavior. What our method also tries to achieve is an empirical test of how typological generalizations may change (or not) when integrating an explicit account of social ecology in our models. While genealogical and geographical bias control has been the only way to account for the social anchoring of language structures in typological models, our methods offer a much wider battery of hitherto largely untested factors, ranging from patterns of language transmission to attitudes and ideologies, which are anchored to social practices across individual domains of interaction.

Acknowledgements

Parts of this article were presented by Kaius Sinnemäki at the Second Workshop of the Nordic Signed Language Corpus Network (NSLCN) at Jyväskylä on 9-10 February 2023, by Francesca Di Garbo, Eri Kashima, Ricardo Napoleão de Souza, and Kaius Sinnemäki at the conference *Naturally Occurring Data in and beyond Linguistic Typology* at Bologna on 18-19 May 2023, and by Eri Kashima, Francesca Di Garbo, and Oona Raatikainen at the *Annual Meeting of the Societas Linguistica Europaea 55* on 8 August 2022. We are grateful to the audiences of these events for their comments, and to two anonymous reviewers for their constructive feedback on an earlier version of this manuscript. This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 805371; PI Kaius Sinnemäki).

Abbreviations

1 = 1st person
3 = 3rd person
DEM = demonstrative
GEN = genitive
OBV = obviative
PL = plural
POSS = possessive
PROX = proximal/proximate
SG = singular

References

Adamou, Evangelia. 2021. *The adaptive bilingual mind: Insights from endangered languages*. Cambridge: Cambridge University Press.

Bell, Allan. 2013. *The guidebook to sociolinguistics*. Oxford: John Wiley & Sons.

Bentz, Christian & Bodo Winter. 2013. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3(1). 1–27. https://doi.org/10.1163/22105832-13030105.

Berdicevskis, Aleksandrs. 2020. Foreigner-directed speech is simpler than native-directed: Evidence from social media. In *Proceedings of the Fourth Workshop on natural language processing and computational social science*, 163–172. Association for Computational Linguistics.

Berthele, Raphael. 2021. The extraordinary ordinary: Re-engineering multilingualism as a natural category. *Language Learning* 71(S1). 80–120. https://doi.org/10.1111/lang.12407.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bickel, Balthasar. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology* 11(1). 239–251. https://doi.org/10.1515/LINGTY.2007.018.

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: In honor of Johanna Nichols*, 415–444. Amsterdam: John Benjamins.

Braunmüller, Kurt Steffen Höder & Karoline Kühl (eds). 2014. *Stability and divergence in language contact: Factors and mechanisms* (Studies in Language Variation 16). Amsterdam: John Benjamins.

Chen, Su-Chiao. 1997. Sociology of language. In Nancy H. Hornberger & David Corson (eds.), *Encyclopedia of language and education: Research methods in language and education* (Encyclopedia of Language and Education, vol. 8), 1–13. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-4535-0_1.

Cysouw, Michael. 2011. Understanding transition probabilities. *Linguistic Typology* 15(2), 415–431. https://doi.org/10.1515/lity.2011.028.

Cysouw, Michael & Jeff Good. 2013. Languoid, doculect and glossonym: Formalizing the notion ‘language’. *Language Documentation & Conservation* 7. 331–359.

Correia, Fabrice & Benjamin Schnieder (eds.). 2012. *Metaphysical grounding: Understanding the structure of reality*. Cambridge: Cambridge University Press.

Croft, William. 2001. *Explaining language change*. Harlow, UK: Pearson Education Limited.

Cronhamn, Sandra. 2018. *Quantifying loanwords*: *A study of borrowability in the Finnish lexicon*. MA thesis, Lund University, Lund.

Di Carlo, Pierpaolo & Jeff Good. 2023. Language contact or linguistic micro-engeeneering? Feature pool, social semiosis and intentional language change in the Cameroonian Grassfields. *Linguistic Typology at the Crossroads* 3(1). 72–125. Special issue on: Language contact and non-convergent change: cases from Africa (edited by Pierpaolo Di Carlo & Pius Akumbu). https://doi.org/10.6092/issn.2785-0943/17231.

Di Garbo, Francesca, Eri Kashima, Ricardo Napoleão de Souza & Kaius Sinnemäki. 2021. Concepts and methods for integrating language typology and sociolinguistics. In Silvia Ballaré & Guglielmo Inglese (eds.), *Tipologia e Sociolinguistica: Verso un approccio integrato allo studio della variazione: Atti del Workshop della Societá Linguistica Italiana 20 settembre 2020*), 143–176. Milano: Officinaventuno. https://doi.org/10.17469/O2105SLI000005.

Di Garbo, Francesca & Ricardo Napoleão de Souza. 2023. A sampling technique for worldwide comparisons of language contact scenarios. *Linguistic Typology* 27(3). 553–589. https://doi.org/10.1515/lingty-2022-0005.

Di Garbo, Francesca & Panagiotis Kapellis 2025. Contact effects in nominal number systems: A world-wide survey. *Studies in Language*. https://doi.org/10.1075/sl.24019.dig.

Dingemanse, Mark, Francisco Torreira & Nick. J. Enfield. 2013. Is “Huh?” a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PLOS ONE* 8(11). e78273. https://doi.org/10.1371/journal.pone.0078273.

Döhler, Christian. 2018. *A grammar of Komnzo*. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.1477799.

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. https://doi.org/10.1038/nature09923.

Fishman, Joshua A. 1965. Who speaks what language to whom and when? *La Linguistique* 1(2). 67–88.

Fleming, Luke. 2011. Name taboos and rigid performativity. *Anthropological Quarterly* 84(1). 141–164.

Floyd, Simeon, Giovanni Rossi & Nick J. Enfield. 2020. A coding scheme for recruitment sequences in interaction. In Simeon Floyd, Giovanni Rossi & Nick Enfield (eds.), *Getting others to do things: A pragmatic typology of recruitments*, 25–50. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.4018372.

Garde, Murray. 2008. *Kun-dangwok*: "clan lects" and *Ausbau* in western Arnhem Land. *International Journal of the Sociology of Language* 191. 141–169. https://doi.org/10.1515/IJSL.2008.027.

Givón, Talmy. 1971. Linguistic colonialism and de-colonialisation: The school system as a tool of oppression. *Ufahamu: A Journal of African Studies* 1(3). 33–49. https://doi.org/10.5070/F713016373.

Griffiths, Devin. 2017. The comparative method and the history of the modern humanities. *History of Humanities* 2(2). http://dx.doi.org/10.1086/693325.

Greenberg, Joseph. 1978. *Universals of human language*. Stanford: Stanford University Press.

Greenhill, Simon J., Xia Hua, Caela F. Welsh, Hilde Schneemann & Lindell Bromham. 2018. Population size and the rate of language evolution: A test across Indo-European, Austronesian, and Bantu languages. *Frontiers in Psychology* 9. 576. https://doi.org/10.3389/fpsyg.2018.00576.

Gumperz, John J. & Robert Wilson. 1971. Convergence and creolization: a case from the Indo-Aryan/Dravidian border in India. In Dell Hymes (ed.), *Pidgnization and creolization of languages*, 151–168. Cambridge: Cambridge University Press.

Guzmán Naranjo, Matías & Gerhard Jäger. 2023. Euclide, the crow, the wolf and the pedestrian: distance metrics for linguistic typology. *Open Research Europe* 3. 104. https://doi.org/10.12688/openreseurope.16141.1.

Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670. https://doi.org/10.1515/lingty-2021-0002.

Haan, Johnson. 2001. *The grammar of Adang: A Papuan language spoken on the island of Alor, East Nusa Tenggara – Indonesia*. Doctoral dissertation, University of Sydney, Sidney. http://hdl.handle.net/2123/6413.

Haig, Geoffrey, Stefan Schnell (eds.). 2023. *Multi-CAST: Multilingual corpus of annotated spoken texts*. Version 2311. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/#veraa (Accessed 2025.12.05).

Haugen, Einar. 1972. *Ecology of language*. Stanford: Stanford University Press.

Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.

Janssen, Dirk, Balthasar Bickel & Fernando Zúñiga. 2006. Randomization tests in language typology. *Linguistic Typology* 10(3). 419–440. https://doi.org/10.1515/LINGTY.2006.013.

Kashima, Eri, Francesca Di Garbo, Oona Raatikainen, Rosnátaly Avelino, Sasha Beck, Anna Berge, Ana Blanco, et al. 2023. *GramAdapt social contact dataset*. Available online at: https://doi.org/10.5281/zenodo.7508054.

Kashima, Eri, Francesca Di Garbo, Ruth Singer & Olesya Khanina & Ruth Singer. 2025. The design principles of a sociolinguistic-typological questionnaire for language contact research. *Language Dynamics and Change* 15(1). 1–103. https://doi.org/10.1163/22105832-bja10035.

King, Brian W. 2019. *Communities of practice in applied language research: A critical introduction*. London: Routledge.

Klamer, Marian. 2012. Papuan-Austronesian language contact: Alorese from an areal perspective. In Nicholas Evans & Marian Klamer (eds). *Melanesian languages on the edge of Asia: Challenges for the 21st century*, 72–108. Honolulu: University of Hawai'i Press.

Kluge, Angela. 2017. *A grammar of Papuan Malay*. Berlin: Language Science Press.

Kusters, Wouter. 2003. *Linguistic complexity: The influence of social change on verbal inflection*. Doctoral dissertation, University of Leiden, Leiden.

Ladd, D. Robert & Roberts, Seán G. & Dediu, Dan. 2015. Correlational studies in typological and historical linguistics. *Annual Review of Linguistics* 1. 221–241. https://doi.org/10.1146/annurev-linguist-030514-124819.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology* 23(3), 533–572. https://doi.org/10.1515/lingty-2019-0025.

Levshina, Natalia. 2022. *Communicative efficiency: Language structure and use*. Cambridge: Cambridge University Press.

Levshina, Natalia, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, et al. 2023. Why we need a gradient approach to word order. *Linguistics* 61(4). 825–883. https://doi.org/10.1515/ling-2021-0098.

Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLOS One* 5(1). e8559. https://doi.org/10.1371/journal.pone.0008559.

Lüpke, Friederike. 2022. *Contact, concord, classification: Noun class systems and categorisation in a multilingual area*. (Paper presented at the seminar of the Helsinki Diversity Linguistics Group, Helsinki, 29 April 2022).

Lüpke, Friederike. 2024. Language, land and languaging in the Atlantic space. In Friederike Lüpke (ed.), *The Oxford guide to the Atlantic languages of West Africa*, 3–15. Oxford: Oxford University Press.

Mansfield, John, Henry Leslie-O'Neill & Haoyi Li. 2023. Dialect differences and linguistic divergence: A crosslinguistic survey of grammatical variation. *Language Dynamics and Change* 13(2). 232–276. https://doi.org/10.1163/22105832-bja10026.

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

Maslova, Elena & Tatiana Nikitina. 2007. Stochastic universals and dynamics of cross-linguistic distributions: The case of alignment types. (Stanford: Stanford University, Unpublished manuscript). http://anothersumma.net/Publications/Ergativity.pdf.

Merlan, Francesca. 1981. Land, language and social identity in Aboriginal Australia. *Mankind* 13(2). 133–148.

Milroy, James & Milroy, Lesley. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics* 21(2). 339–384. https://doi.org/10.1017/S0022226700010306.

Moro, Francesca R. 2018. The plural word *hire* in Alorese: Contact-induced change from neighboring Alor-Pantar languages. *Oceanic Linguistics* 57(1). 177–198. https://doi.org/10.1353/ol.2018.0006.

Moro, Francesca R. 2021. Multilingualism in eastern Indonesia: linguistic evidence of a shift from symmetric to asymmetric multilingualism. *International Journal of Bilingualism* 25(4). 1102–1119. https://doi.org/10.1177/13670069211023134.

Moro, Francesca & Yunus Sulistyono. forthcoming. The Alorese and the Adang in eastern Indonesia. In Francesca Di Garbo & Eri Kashima & Kaius Sinnemäki (eds.), *Social foundations of language contact: A comparative survey*. Berlin: Language Science Press.

Muysken, Pieter. 2013. Language contact outcomes as the result of bilingual optimization strategies. *Bilingualism: Language and Cognition* 16(4). 709–730. https://doi.org/10.1017/S1366728912000727.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2017. *Historical sociolinguistics: Language change in Tudor and Stuart England*. 2nd edn. London: Routledge.

Nishiyama, Kunio & Herman Kelen. 2007. *A grammar of Lamaholot, eastern Indonesia: The morphology and syntax of the Lewoingu dialect.* München: Lincom.

Pakendorf, Brigitte Nina Dobrushina & Olesya Khanina. 2021. A typology of small-scale multilingualism. *International Journal of Bilingualism* 25(4). 835–859. https://doi.org/10.1177/13670069211023137.

Ross, Malcolm. 2013. Diagnosing contact processes from their outcomes: The importance of life stages. *Journal of Language Contact* 6(1). 5–47. https://doi.org/10.1163/19552629-006001002.

Rossi, Giovanni, Simeon Floyd & Nick J. Enfield. 2020. Recruitments and pragmatic typology. In Simeon Floyd, Giovanni Rossi & Nick J. Enfield (eds.), *Getting others to do things: A pragmatic typology of recruitments*, 1–23. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.4018370.

Seifart, Frank Ludger Paschen & Matthew Stave (eds.). 2024. *Language documentation reference corpus (DoReCo)* 2.0. Berlin & Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Available online at: https://doi.org/10.34847/nkl.7cbfq779 (Accessed 2025.12.05).

Silverstein, Michael. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & Communication* 23(3). 193–229. https://doi.org/10.1016/S0271-5309(03)00013-2.

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

Sinnemäki, Kaius. 2010. Word order in zero-marking languages. *Studies in Language* 34(4). 869–912. https://doi.org/10.1075/sl.34.4.04sin.

Sinnemäki, Kaius. 2014. Cognitive processing, language typology, and variation. *WIREs Cognitive Science* 5(4). 477–487. https://doi.org/10.1002/wcs.1294.

Sinnemäki, Kaius & Francesca Di Garbo. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology* 9. 1141. https://doi.org/10.3389/fpsyg.2018.01141.

Sinnemäki, Kaius. 2020. Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach. *Journal of Historical Sociolinguistics* 6(2). 1–39. https://doi.org/10.1515/jhsl-2019-1010.

Sinnemäki, Kaius & Ahola, Noora. 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. https://doi.org/10.1111/1467-968X.12284.

Sinnemäki, Kaius, Francesca Di Garbo, Mark Ellison & Ricardo Napoleão de Souza. 2024. A typological approach to language change in contact situations. *Diachronica* 41(3). 379–413. https://doi.org/10.1075/dia.23029.sin.

Sinnemäki, Kaius. (in revision). On the "socio" in sociolinguistic typology: A review.

Sinnemäki, Kaius & Eri Kashima (forthcoming). Comparative historical sociolinguistics. In Terttu Nevalainen, Bridget Drinka, & Gijsbert J. Rutten (eds.), *Handbook of historical sociolinguistics*. Berlin: De Gruyter Mouton.

Suokhrie, Kelhouvinuo. 2016. Clans and clanlectal contact: Variation and change in Angami. *Asia-Pacific Language Variation* 2(2). 188–214. https://doi.org/10.1075/aplv.2.2.04suo.

Sulistyono, Yunus. 2022. *A history of Alorese (Austronesian) Combining linguistic and oral history*. Doctoral dissertation, University of Leiden, Leiden.

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

Tomasello, Michael. 2010. *Origins of human communication*. Cambridge, MA: MIT Press.

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

Wenger, Etienne. 1998. *Communities of practice: Learning, meaning, and identity*. Cambridge: Cambridge University Press.

Winter, Bodo, & Grice, Martine. 2021. Independence and generalizability in linguistics. *Linguistics 59*(5). 1251–1277. https://doi.org/10.1515/ling-2019-0049.

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 (Open Handbooks in Linguistics)*, 631–642. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0061.

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

Zeman, Daniel et al. 2023. *Universal Dependencies 2.13*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of

Mathematics and Physics, Charles University. Available online at: http://hdl.handle.net/11234/1-5287.

CONTACT

francesca.di-garbo@univ-amu.fr
kaius.sinnemaki@helsinki.fi
eri.kashima@helsinki.fi or eri.kashima@amu.edu.au