Dialect differences and grammatical divergence: A cross-linguistic survey

John Mansfield, Henry Leslie-O'Neil and Haoyi Li

Abstract
This article presents a new type of comparative linguistic survey, analysing on a database of 1181 linguistic and sociolinguistic variables drawn from 42 genealogically and geographically diverse languages. We focus in particular on grammatical variables, and whether they do or don't differentiate geographic dialects. We identify three main structural types of grammatical variable: FORM, ORDER and OMISSION, and find that in situations of close contact between dialect groups, form variables are more likely to differentiate dialects than the other two types. Order and omission variables usually only differentiate dialects that have minimal contact. Our findings suggest that signalling of group identity may have a role in the divergence of grammars, though this affects some dimensions of grammar more than others.

Keywords
grammatical variation, linguistic divergence, dialectology, sociolinguistics, language contact

1. Introduction

Language simultaneously expresses semantic content, and signals aspects of social identity. Comparative linguistics has explored in great detail how linguistic form maps to semantic content, but there has been little comparative research on the relationship of language structure to social identity. In this study we address this gap using a new type of data: a survey of linguistic and sociolinguistic variables from a wide range of language families.

Our interest in social signalling lies particularly in the role it plays in fomenting differences between language varieties. Under one model of linguistic differentiation, social groups separate and progressively lose contact, which allows their respective linguistic codes to gradually become more different from one another. Linguistic differentiation is a function of time spent apart. But in recent years another type of differentiation has come to light, sometimes known as 'linguistic divergence'. Studies of small-scale multilingualism have shown that social groups may live in very close interaction, while nonetheless carefully policing their language borders, deploying social norms and interactional etiquette to maintain or enhance the distinctness of their linguistic codes (e.g. Di Carlo 2018; Evans 2019; Epps 2020). Even the differences between closely-related dialects may be carefully cultivated to construct distinctive group affiliations (Morphy 1977; Stanford 2009; Vaughan 2018). Evolutionary theorists conjecture that this sort of lectal differentiation may play a role in controlling access to local networks of mutual assistance (Nettle & Dunbar 1997; Dunbar

2003), and artificial language experiments have lent some support to this, by successfully simulating divergence of lects under social pressures (e.g. Roberts 2010; Sneller & Roberts 2018; Lai et al. 2020). Furthermore, a large-scale study of vocabulary differentiation supports the concept of 'punctuational bursts', that is, language varieties changing their vocabulary more rapidly as part of the process of social-group fission (Atkinson et al. 2008).

But there may also be purely cognitive, as opposed to socio-cultural, factors at work. In a study of bilingual production, where the two languages share a large number of similar forms, speakers exhibited a bias against those forms of ambiguous provenance, which suggests that bilingual processing could drive lexicons apart (Ellison & Miceli 2017). Although these studies encompass both socio-cultural motivations, and unconscious cognitive biases, they all point to the potential for language varieties to diverge because of the social interaction between groups. We can therefore define 'linguistic divergence' as differentiation that is driven by language contact, rather than the absence of contact.

Linguistic divergence raises a series of important questions for the study of language change, such as: What types of linguistic structure are affected? What parameters of social interaction promote divergence? What types of group relations? How might divergence be incorporated into our models of language phylogeny?

In this study we focus on one small part of the puzzle. We study pairs or clusters of dialects, that is, closely related language varieties that are associated with distinctive geographic territories, drawing our data from 42 reference grammars of geographically and genetically dispersed languages. We extract data on *grammatical variables* in these languages, which are grammatical meanings or functions that can be expressed in more than way (for example, in English future tense can be expressed by both *will* and *gonna*). For each of these grammatical variables, the crucial question is whether it distinguishes dialects, or cuts across dialects. We apply a simple structural typology of grammatical variables, and ask whether dialects are more likely to be differentiated by some types of grammatical variables, rather than others. We also estimate degrees of social contact between dialect groups, as this appears to play an important role in the patterning of structural types.

We identify three basic structural types of grammatical variable: FORM variables, which involve distinct grammatical markers appearing in the same linear position, as in (1); ORDER variables, which involve the same linguistic elements in different linear orders, as in (2); and OMISSION variables, which involve the presence/absence of a grammatical marker, as in (3). Form and order variables correspond to the paradigmatic and syntagmatic dimensions of language, respectively (Bloomfield 1933; Saussure 1959), while omission variables are related to notions such as redundancy and underspecification.

(1) FORM VARIABLE (Kugu Nganhcara comitative)

thuli-ra ~ *thuli-nta*
woomera-COM
'with a woomera' (Smith & Johnson 2000: 393)

(2) ORDER VARIABLE (Komnzo adjectival attribution)

zagr *karfo* *~ karfo zagr*
distant village
'distant village' (Döhler 2018: 89)

(3) OMISSION VARIABLE (Tundra Nenets comparison)

t'uku° *pəni°* *taki°* *pəne-xəd°* *səwa(-rka)*
this coat that coat.ABL good-COMP
'This coat is better than that one.' (Nikolaeva 2014: 174)

Our main finding is that grammatical form variables frequently differentiate dialects, and they are equally likely to differentiate dialects whether they are in close contact or not (in example 1, the variants are associated with nearby Uwanh and Iyanh dialects respectively). Order and omission variables, by contrast, only rarely differentiate close-contact dialects, though they frequently differentiate dialects that are relatively distant from one another.

The remainder of this article runs as follows. Section §2 establishes our framework for conceptualising linguistic variables, dialect groups and social signalling. Section §3 introduces the database used for this study. Section §4 outlines our three basic structural types of grammatical variable, with informal observations on how these relate to sociolinguistics and dialect relations. Section §5 provides a formal quantitative analysis, lending support to the hypothesis that the structural types function differently with respect to dialect contact and differentiation. Section §6 summarises our findings and discusses implications for further research.

2. Dialect differences, social contact and social-indexicality

In this study we consider language varieties to be in a 'dialectal' relationship whenever they share the vast majority of their grammar, phonology and lexicon, but nonetheless are different enough to be recognised as distinct varieties. We call such relations 'dialectal' when they are based on geography (e.g. different towns, provinces or regions), as opposed to other types of language variety associated with socio-economic groups, subcultures, formal registers etc. Note that this approach to dialects is independent of ethno-linguistic naming practices. For example, there is a dialectal relationship between varieties spoken on the Aguaytía and San Alejandro rivers in Peru, both of which are known by the label Kakataibo (Zariquiey 2011; Zariquiey 2018: 3). But there is also a dialectal relationship between Emmi and Mendhe (Ford 1998), two very similar lects associated with neighbouring clan estates in northern Australia, which do not share an ethno-linguistic label.

The concept of 'dialect' has often been used for unwritten regional varieties in relation to written, supra-regional 'standard' varieties (e.g. Haugen 1988; Chambers & Trudgill 1998; Abraham 2006). But this approach is only relevant in those circumstances where there is a supra-local standard, such as a national language. The current study is broadly concerned with human languages, most of which have no supra-local standard form. Therefore most of the dialectal relations in this study are between pairs of closely-related, regional varieties, usually without any political hegemony of one over the other.

Dialectal differences develop through the separation of social groups into distinct geographic territories, with concomitant separation of an erstwhile shared language variety into two distinct varieties. Because people tend to interact with those who live near them, geography plays a major role in the development of sociolinguistic groupings (Paul 1888: 23ff.; Trudgill 1986: 39). For example, from the tenth century until the sixteenth century there was a fairly discrete and integrated community living on the island of Jersey, speaking a shared variety of Norman French. In the sixteenth century, forty families from Jersey moved to the smaller island of Sark, leading to the subsequent divergence of a Sark dialect from the Jersey dialect (Liddicoat 1994: 6).

Language differentiation does not always follow a smooth trajectory of separation. Dialects may remain in close contact for hundreds or thousands of years, remaining similar because they have remained in continual contact and shared many linguistic innovations. There are also instances where dialects are in a process of convergence, rather than divergence (Trudgill 1986). More generally, we must recognise that language histories are not always made up of neat iterative splits (e.g. Garrett 2006; François 2014), and the formal similarity involved in dialectal relationships can arise from any type of relatively recent social contact, with concomitant language contact. But irrespective of these diverse histories, dialects have an important role in the larger-scale process of language diversification. In the development of distinct languages, where eventually the grammar, phonology and lexicon drift far apart, there must be some early stage at which the differences are more subtle, and it is this stage which we conceptualise as a dialectal relationship. In this study we focus on dialect relations as a key early stage in linguistic differentiation, and our main findings are synchronic observations of how social contact patterns with certain types of grammatical differentiation. But we will also use these findings to consider dialectal relations as one stage in a larger diachronic process, making predictions about likely trajectories of change (§6).

2.1. Variables and dialects

We conceptualise linguistic variation in terms of VARIABLES, where a variable involves two or more expressions that have the same semantic content (Weinreich et al. 1968: 159). Given a pair of variant expressions, $x_1 \sim x_2$, an individual language user, at a given point of time, has a particular probability of selecting one variant or the other. At the extremes are individuals who categorically use just one variant or the other.

We assume that individuals form dialect groups, which are groups of individuals who interact with each other more than they interact with those outside the group (Croft 2000: 20). However there is also some degree of interaction between individuals in different groups. Figure 1(a) shows two dialect groups, one above and one below. Within each group there are dense social connections (solid lines), but there are also some social connections (dotted lines) running between dialect groups. Each individual is shaded in greyscale, representing their probability of using variants $x_1 \sim x_2$. For variable (a) we see that most individuals use both variants, and the two groups have similar distributions. We can say that this is a 'intra-group' or 'non-dialectal' variable. For variable (b) we see the same pair of dialectal groups, but here variant selection is strongly biased towards x_1 in the top group, and x_2 in the bottom

group. We can say that this is a 'dialect variable' – noting that it does not require a categorical split between dialect groups, but only that there is a notable difference between dialect groups with respect to the variable.

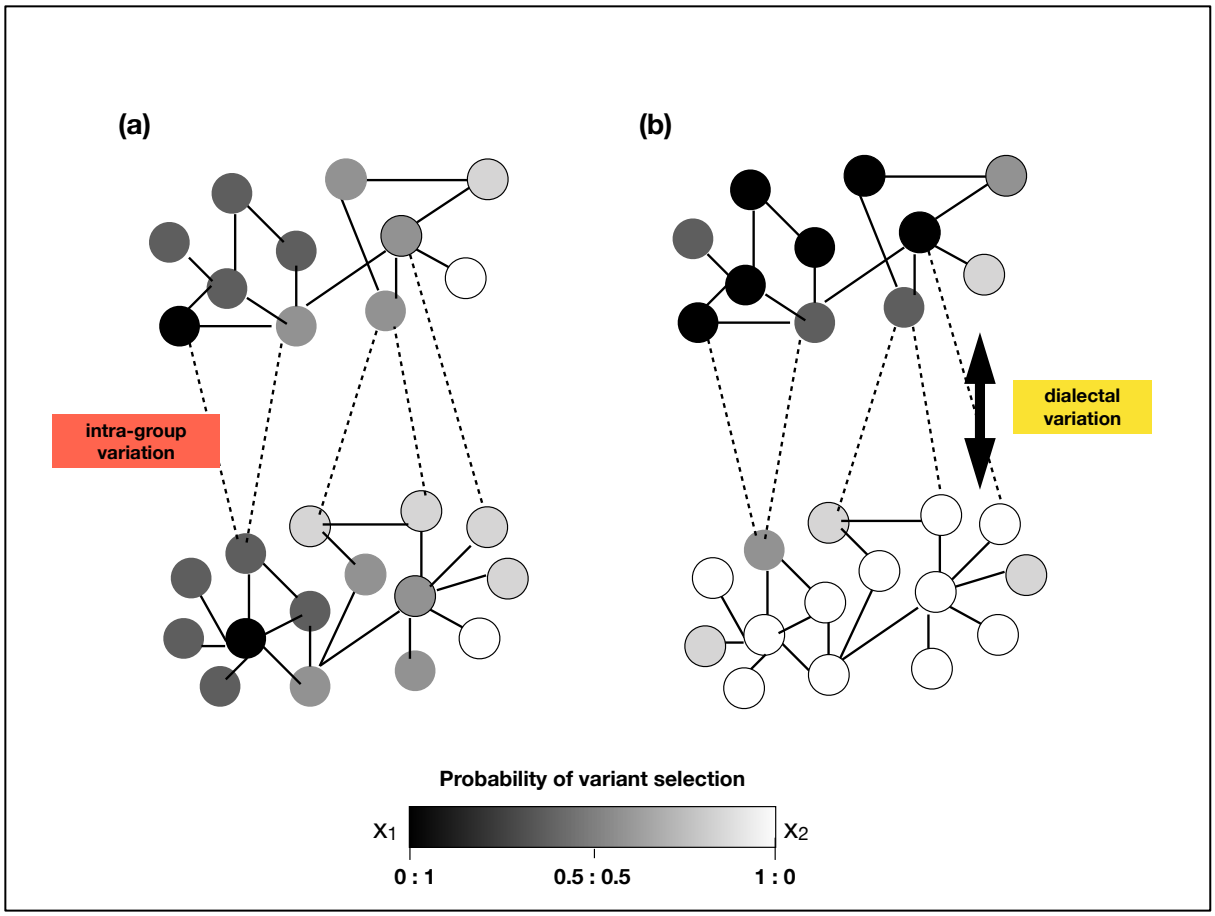

Figure 1. Linguistic variables and dialect groups: (a) Where the selection of variants $x_1 \sim x_2$ has a similar distribution in two dialect groups, we call this 'intra-group' or 'non-dialectal' variation; (b) Where the selection of variants $x_1 \sim x_2$ aligns with dialect groups, we call this 'dialectal' variation.

2.2. Social-indexicality and language structure

In Figure 1(b) above, we use dashed lines to represent interaction between individuals in different dialect groups. Where this interaction is substantial, it is plausible that dialect variation provides cues about group affiliation. Because there is substantial interaction between the groups, individuals would have some exposure to both dialectal variants, and could form conscious or unconscious associations between variants and group identity. Recent studies have suggested that group interaction of this type can result in linguistic divergence, that is, the differentiation of language varieties is facilitated by language contact, with individuals associating distinct expressions with distinct language varieties to which they are exposed (e.g. Di Carlo 2018; Evans 2019; Epps 2020). Following sociolinguistic literature, we use the term 'social-indexical' in reference to these types of variables – that is, linguistic cues that function to indicate different ways of speaking associated with different social groups (Agha 2003; Silverstein 2003; Eckert 2008; Eckert 2019).

However, we must also consider that dialectal differences may develop without any social-indexicality. This is especially likely once dialect groups have such reduced contact that individuals would not have sufficient exposure to both variants, and the group identities would be less relevant to managing social relations. When groups are socially separated, over time they may develop dialectal differences independent from the social-indexicality effect; furthermore, such innovations have less chance of spreading between the groups, due to lack of contact.

Previous work in sociolinguistics has considered whether certain types of language structure are more or less amenable to social-indexicality. There may be cognitive or communicative constraints that make it easier to associate certain types of formal distinction with social identity. Such notions have been discussed in the sociolinguistics literature using various terms, such as 'marker vs indicator' (Labov 1972) 'metapragmatic awareness' (Silverstein 1981), 'pragmatic salience' (Errington 1985) and 'sociolinguistic salience' (Kerswill & Williams 2002; Rácz 2013; Levon & Buchstaller 2015). One enduring idea has been that 'surface' linguistic forms are more capable of social-indexicality than 'deep' linguistic structure (e.g. Labov 1993; Hinskens 1998; see also Eckert 2019). Thus phonology and lexicon are more social-indexical, while morphology is less so, and syntax is the least social-indexical of all (Romaine 1981; Cheshire 1987; Dediu et al. 2013: 311). Similarly, studies in comparative linguistics suggest that communities in contact tend to differentiate themselves using the forms of lexemes or grammatical markers, while unconsciously converging in their morphosyntactic structures (Gumperz & Wilson 1971; Grace 1981; Ross 1996; Ross 2001). At the same time, it can be difficult to determine which variables should be counted as surface forms, and which apparent 'surface forms' actually reflect variation in underlying structure (Meyerhoff & Walker 2012). In this study we follow the lead of these earlier works in attempting to show that some dimensions of language have greater potential for social-indexicality than others. But rather than applying a surface vs depth model, which depends on specific analyses of structural layers, we instead focus on paradigmatic versus syntagmatic dimensions of surface structure (see §4 below), since these can be applied in a relatively theory-neutral way based on linguistic documentation.

3. Collating a cross-linguistic sample of dialect differences

Many typological databases aim to distil the information in reference grammars, in order to summarise grammatical difference. Linguistic variation is a kind of noise that such databases must filter out. In this study we take the opposite approach, specifically targeting whatever reference grammars report to be variable (see also Di Garbo et al. 2021). We extracted data from reference grammars of 42 languages (see map in Figure 2), representing 28 different language families, and all inhabited continents. Although this is only a small sample of the world's linguistic diversity, it is uniquely systematic and wide-ranging within the nascent field of comparative sociolinguistics. All languages in the current sample are spoken languages, though we aim to include signed languages in future work.

Variables were added to the database by searching reference grammars for mentions of variation, using a combination of keyword searches and reading (see Supplementary Information for further details). A grammatical variable was coded wherever the text reports more than one way of expressing the same grammatical meaning or function. Grammatical meanings are relatively abstract categories, such as future, negation, continuous aspect, first-person, directionality; or functions such as focus, subordination or transitivity (Lehmann 1995; Hopper and Traugott 2003; Boye and Harder 2012). This method yielded 1181 grammatical variables, for which basic structural type, and dialectal status, were entered into a spreadsheet, then transformed into an R datatable (see examples in Table 2 below). Most grammars also mention a range of phonological and lexical variables, which we have noted for further research but are not included in this study.

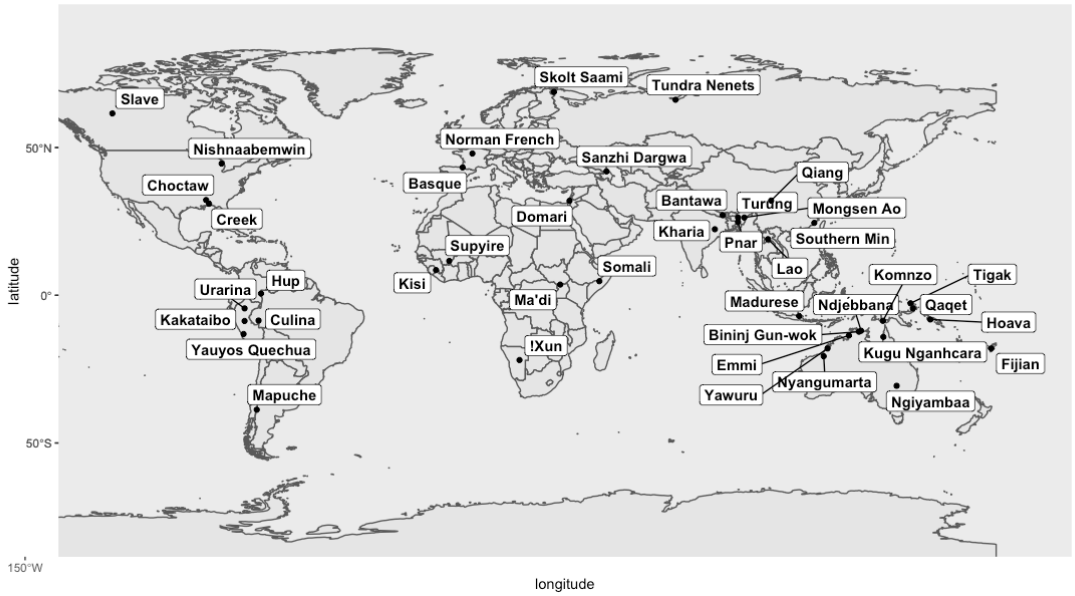

Figure 2. Languages sampled in our database.

Our data was selected to represent a range of diverse languages and social situations, but the sample is not balanced either by language family or region. Most language families are represented by a single language, but a few (mostly larger families) have multiple languages. Furthermore, the number of datapoints contributed by each language family varies widely, since some grammars yielded more variables than others. Figure 3 shows the number of grammatical variables contributed by each language family, and how many of these are dialectal variables. Most families contributed between 10 and 50 data points, while others contributed 100 or more. Austronesian and Sino-Tibetan contributed more data because our sample includes several grammars, representing distinct branches of these families. But in the case of Athapaskan, and to a lesser extent Basque, we have sampled just one grammar from each of these families (Athapaskan: Rice 1989; Basque: Hualde & Urbina 2003), but these two sources were unusually rich in grammatical variables. We accept this imbalance because it allows us to capture all the information provided by the reference grammars, while the statistical problem can be adequately managed by using a mixed-effects regression model

with language families as group effects (§4.4). For 26 of the 28 language families, at least one of the grammatical variables was reported to be dialectal.[1]

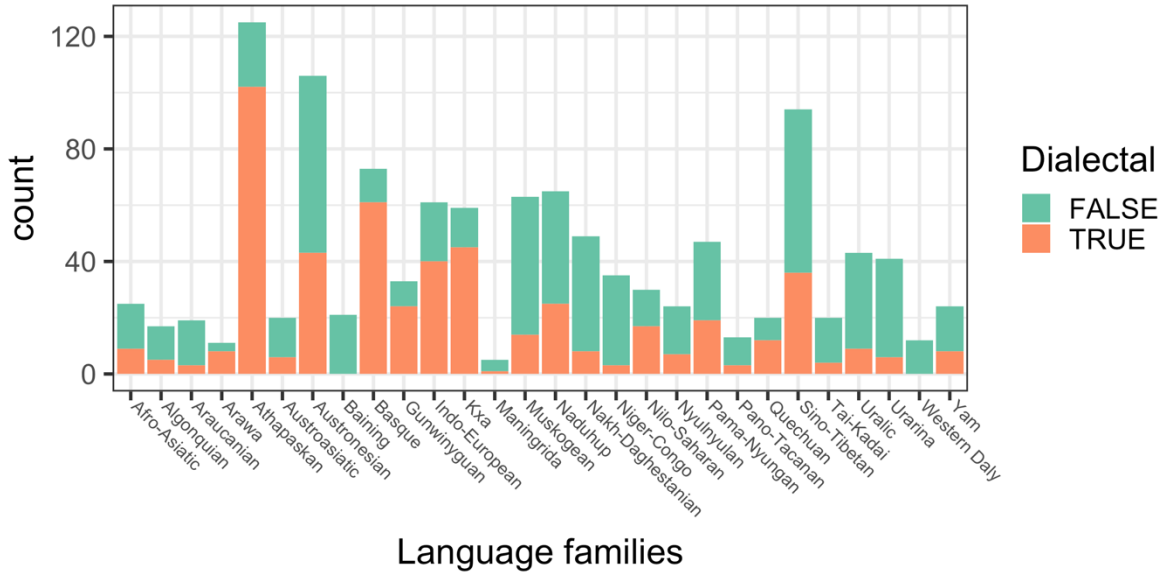

Figure 3. Number of grammatical variables per language family, dialectal and non-dialectal.

We annotate variables wherever the source presents expressions as having the same meaning, but we do not attempt to further investigate whether these expressions have exactly the same connotations or truth-conditional semantics. Sociolinguists working on grammatical variation have long recognised this as a difficult problem (cf. Lavandera 1978; Romaine 1981; Cheshire 1987 inter alia). We take an onomasiological approach, i.e. using meaning as a starting point, rather than form. Consequently a grammatical variable is annotated wherever two expressions can convey the same grammatical meaning; but one or both of these expressions may be also capable of expressing other meanings. For example, the meaning FUTURE may be expressed alternately by a specifically future tense marker, or by a variant that spans both future and present (i.e. non-past) meanings. This is an important point to which we return below (§4.1). We also note that our data coding is not directly comparable to some other studies of grammatical change (e.g. Greenhill et al. 2017; Matsumae et al. 2021), which use features from the World Atlas of Linguistic Structures (Dryer & Haspelmath 2005). Only a subset of our grammatical variables correspond to features coded in the World Atlas.

Reference grammars usually attest variation in a succinct, impressionistic form, glossing over the nuances of variant distributions. In our Figure 1(b) schema, we noted that variants may each have some usage in each group, while nonetheless making a stochastic group distinction. This is mirrored in reference grammars, which sometimes describe

[1] The two exceptions are Baining (Hellwig 2019) and Western Daly (Ford 1998). These attest lexical and phonological differences between dialects, and repors grammatical variables which are non-dialectal, but they do not attest any variables that are both dialectal and grammatical.

categorical dialectal variables, and sometimes note that one variant is 'more common' in one dialect than another. The coding used in this study represents both of these situations as dialect variables, without distinguishing categorical from stochastic types.

3.1. Limitations of the method

An important limitation of our method is that some reference grammars pay closer attention to dialectology than others, meaning that our primary data is partial and approximate. A grammar writer may mistakenly report a dialectal variable, based on what is actually intra-group variation. Alternatively, what is presented as a non-dialectal or 'free variation' may on closer inspection turn out to be dialectal. We must therefore assume that there is a certain degree of noise in our data sources.

Another limitation of the data is the difficulty of coding up the grammars in a fully reproducible way. Coding was performed by all three authors of this article, with most grammars being coded by multiple authors to improve consistency (see Supplementary Materials for details of the coding method and intercoder reliability). We found that our coding of structural types and dialectal status were quite consistent, but it was difficult to achieve consistency on exactly how many variables are identified in a given section of a reference grammar. We therefore do not treat the *number* of variables as an interpretable finding, instead focusing on patterns in structural types and dialectal status. Although our database cannot claim to be either comprehensive or fully reproducible, we have no reason to expect that these limitations should invalidate the findings presented in this study. Our methodological limitations would invalidate the findings if there were systematic inaccuracies in whether structural types are identified as dialectal or non-dialectal, but we do not have any reason to expect such systematic errors.

Compared to the reference grammars used in this study, dedicated sociolinguistic studies could provide more detailed information about specific variables and their (stochastic) group associations. But variationist sociolinguistics does not offer a large enough sample of variables from diverse languages, as the field is still heavily focused on a small number of politically dominant, cosmopolitan languages (Stanford 2016; Mansfield & Stanford 2017). We preferred reference grammars because they provide a more diverse linguistic sample. But another important advantage is that grammars include information on both dialectal and non-dialectal variables, which is crucial to identifying which types of structure are more or less likely to differentiate dialects.

3.2. Coding dialect contact

As well as coding individual linguistic variables, for each reference grammar we also coded degrees of social contact between dialect groups. Reference grammars provide information on social relations, either directly by reporting on social interaction, or indirectly in comments mutual intelligibility of dialects, geographic proximity etc. We used this information to create a rubric for assigning dialectal relations to three degrees of social distance: Close, Medium, Distant (see Supplementary Materials for details). This is an admittedly coarse and informal measure, which does not capture the nuances of social

relations among groups. Nor does it capture diachronic dynamics, with social relations changing from one historical period to another. Nonetheless, it was important to parameterise social contact in our data since the dialect relations reported in the grammars clearly encompassed very different degrees of contact, as illustrated by the following examples.

The Kugu Nganhcara grammar (Smith & Johnson 2000) reports the very closest type of dialect relations. The language as a whole is reported to have about 300 speakers, but within this population speakers identify with six different patriclans, each of which is associated with distinct geographic territory, and has its own dialect or 'clan lect' (Smith & Johnson 2000: 358). However, rather than living separate lives on their separate territories, people from each clan group are highly mobile, and often live intermingled in the same residential groups, for example when jointly exploiting natural resources. The mingling of residential groups is also ensured by clan exogamy (marriage between people from different clans). Thus there is extensive interaction between speakers of different clan lects, and we code Kugu Nganhcara dialect relations as Close.

An intermediate level of contact is found in the grammar of Channel Island French (Liddicoat 1994), which focuses on dialects from the islands of Jersey and Sark. As mentioned above, the Sark community split off from Jersey in the sixteenth century. Both dialect groups have had predominantly agricultural livelihoods since then, with social interaction organised around local villages and their markets. This implies a lower level of contact between the two dialect groups. On the other hand, the distance between the islands is small and easily navigable (about 30km), and the agricultural communities have been involved in significant cross-channel trade. We assigned this dialect relation a Medium contact value.

A Distant dialect relation is found in Somali (Saeed 1999), a language spoken by several million people across a large region. Northern dialects are spoken by pastoralists living on relatively arid country, and southern dialects are spoken by agriculturalists living in a river delta some hundreds of kilometres to the south. Mutual intelligibility is asymmetrical, with southerners able to use northern dialect as a lingua franca, but northerners being less familiar with the southern dialect.

Note that for most grammars (e.g. Kugu Nganhcara), we coded the same degree of social distance for all dialect relations. But for other grammars (e.g. !Xun), some dialect relations were judged to be more distant than others. This is also the case in Hup (Epps 2008), where the grammar reports a generally high level of mutual intelligibility, and notes that the main social groups, patrilineal clans, live alongside each other in shared villages. On this basis the central and eastern dialect areas of Hup are coded as a Close dialect relationship. However the western dialect speakers have less interaction with the central and eastern groups, and the central/eastern speakers say the western dialect is 'hard to understand' (Epps 2008: 13). On this basis, we assigned a Distant relationship between western dialect and the other two.

Table 3 shows the coding of some example variables. Most of the examples shown here are dialectal variables, but Kugu Nganhcara SOV ~ SVO, and Nishnaaabemwin nominal conjunction, are examples of cross-cutting variables (Dialects = NA). There are two dialectal

variables from Hup, but one of these is between the Close villages, while the other is between the Distant western dialect compared to other areas. (The two Hup variables are also examples of stochastic dialect variables: in each instance, one of the dialect groups is reported to use both variants.) The coding of the Type column will be explained in the following sections.

Table 2. Examples of grammatical variable coding
(see Supplementary Information for further details and references)

LANGUAGE	MEANING	VAR1	VAR2	TYPE	DIALECTS	CONTACT
Kugu Nganhcara	Comitative	N-ra	N-nta	Form	Uwanh ~ Iyanh	Close
Kugu Nganhcara	Pronoun 1.PL.EXCL	ŋaɲca	ŋana	Form	Muminh ~ Iyanh	Close
Kugu Nganhcara	Transitive clause	SVO	SOV	Order	NA	Close
Basque	Future participle (stem ending /n, l/)	V.PTCPL-ko	V.PTCPL-en	Form	western ~ eastern	Distant
Hup	Adj-INTNS	Adj-Vcap	Adj-icap	Form	Barriera~ Tat Deh	Close
Hup	V-INTNS	V-tubud	V-túud	Form	others ~ western	Distant
Nishnaabemwin	Nominal conjunction	NP conj NP	NP NP	Omit	NA	Medium
Nishnaabemwin	Plural	N-o:g	N-ag	Form	NA	Medium

4. Structural types and dialectal status

Our grammatical variables were coded into structural types, with categories developed iteratively as coding proceeded. Three main types were identified:[2]

(a) FORM: Variants have the same structure, but are distinguished by the form of a grammatical marker (either affix, clitic or function word);

(b) ORDER: Variants use the same lexical and grammatical elements, but are distinguished by linear ordering;

(c) OMISSION: Variants are identical except that a grammatical marker is present in one but absent in the other.

The following subsections describe each type in turn, and make general observations about their dialectal or non-dialectal status.

4.1. Form variables

A FORM variable is where variant expressions of a grammatical meaning are distinguished by the form of the grammatical element, but in other respects the construction is the same. A well-studied example in English involves negative predicates, which vary in the form of the negative auxiliary/copula, e.g. *she isn't home ~ she ain't home*. This a social-indexical variable, marking social class, stance and style (Levinson 1988; Cheshire et al. 2005). English has other well-known grammatical variables that are also form variables, and also

[2] A small residue group of variables (8% of the total) exhibit a mixture of the criteria for the three main types. See Supplementary Materials. These are excluded from the analysis below.

have strong social-indexicality. These include 'negative concord', involving paradigmatic contrast between negative determiners *any ~ no* (Wolfram 1969), and the verbal progressive suffix *-ing ~ -in* (Campbell-Kibler 2010). Latin American Spanish offers another well-studied example in the expression of second-person singular subject, where the *voseo* phenomenon involves distinctive 2SG markers both in free pronouns and in verbal suffixes. In some areas *voseo* is a recognised marker of regional dialects, for example in Colombia (Collazos 2015: 10ff.; Fernández Acosta 2020):

(4) *Colombian Spanish*

a. (tu) com-es (e.g. Cordoba dialect)

b. (vos) com-és (e.g. Antioquia dialect)
2SG.Seat-2SG.S
'You eat.'

A particularly flamboyant example of dialectal form variation is in Bininj Gun-wok, where certain verbal prefixes index patrilineal clan heritage, which affords rights to territorial estates (Garde 2008). What makes this example so striking is that the prefixes do not carry any semantic content: they are semantically vacuous 'fillers', purely sociolinguistic markers used to stake out affiliations to clan and land:

(5) *Bininj Gun-wok*

a. yi-njarra-kinje-men (Djordi clan)

b. yi-bayid-kinje-men (Kurulk clan)

c. yi-buk-kinje-men (Mok clan)
2SG-CLAN.INDEX-cook-IMP
'You cook it!' (Garde 2008: 150–154)

Form variables may emerge either from sound changes, or grammaticalisation. A phonologically induced example can be seen in the American English 1SG.FUT auxiliary, where African-American dialects have innovated *I'm'a* VP, marking a point of differentiation from other dialects *I'm gəna* VP. Here phonological erosion has been applied differently in different dialects. Although such variables have a phonological dimension, we still treat them as grammatical variables wherever the sound change appears to be specific to a grammatical marker, as opposed to being a regular sound change.

Form variation via grammaticalisation paths can be seen in the second-person plural pronoun in English dialects, which may take the form *youse* (e.g. Australian) or *y'all* (e.g. southern USA), exhibiting different grammaticalisation paths in the development of the pluralising suffix.

Form variables are the most frequent type of grammatical variable in our data, accounting for 56% (N=664) of all variables annotated. There is at least one grammatical form variable reported in each of the 42 languages. Form variables are also the type in which the highest proportion are dialectal, with 58% (N=385) of form variables being dialectal.

Form variation of grammatical markers therefore appears to be a cross-linguistically frequent type of dialectal differentiation.

The types of grammatical markers involved in form variables include affixes, clitics and function words, and encompass a wide range of grammatical meanings and functions. Examples (6–8) illustrate pronominal variation in free pronouns, clitics and affixes respectively. Here and below, for each variable we indicate whether it is reported to have a dialectal association or not.

(6) *Fijian personal pronoun*

koya ~ ʔea	(Standard/Bau dialect ~ Boumaa dialect)
'him'	(Dixon 1988: 54)

(7) *Kharia pronominal suffix for irrealis middle verbs*

V=na=iŋ ~ V=na=ŋ	(no dialectal association)
V=MID.IRR=1sg	(Peterson 2010: 249)

(8) *Bininj Gun-wok pronominal prefix*

karri-V ~ yirri-V	(Kunwinjku dialect ~ Gundedjnjenghmi dialect)
1.INCL.AUG-V	(Evans 2003: 20)

Examples (9, 10) illustrate variation in case markers and prepositions.

(9) *Basque case suffix*

N-etarik ~ N-tik	(eastern region ~ other regions)
N-ABL	(Hualde & Urbina 2003: 185)

(10) *!Xun locative post-position*

N sí ~ N ǹǃŋ́	(E1 vs other dialects)
N LOC	(Heine & König 2015: 187)

Examples (11–13) illustrate variation in TAM, quantification and assent.

(11) *Choctaw imperative suffix*

V-tʃa ~ V-oː	(no dialectal association)
V-IMP	(Broadwell 2006: 194)

(12) *Emmi quantifier*

dawal ~ pakwuc	(no dialectal association)
'many'	(Ford 1998: 194)

(13) *Urarina assent*

ajara ~ ẽehe	(Asna dialect ~ other dialects)
'yes'	(Olawsky 2006: 882)

In summary, our data suggests that dialectal form variables can occur in any grammatical function, at any constituency level (affix, clitic or phrasal). Although a much larger database

would be required to investigate whether dialectal differentiation is more or less likely in these semantic and structural types, our data does not suggest any obvious constraints on dialectal association of form variables.

As mentioned above, we define grammatical variables as two ways of expressing a grammatical meaning/function, even if these two expressions may themselves have differences of functional range (e.g. a specific FUT marker vs a more general NON-PAST marker). When such markers distinguish dialects, this implies that the dialects are *not* grammatically isomorphic. It would therefore be misleading to say that such dialects differ only on their 'surface' forms, since the 'underlying' structure of grammatical features is also different. Investigation of such non-isomorphisms may reveal important patterns of grammatical divergence, however this is beyond the scope of the current study.

4.2. Order variables

An ORDER variable is where two variant expressions are composed of the same combination of forms, but the linear ordering is different.[3] Order variables may involve positioning of a grammatical marker, or re-ordering of lexical elements, without a change of meaning. For example, Spanish object clitics may be positioned either after an infinitive verb or before the finite verb:

(14) *Spanish*

no	puede	manejar=los	~	no los=puede manejar
NEG	can.3SG.PRS	manage=3PL.M.OBJ		

'She can't manage them.' (Schwenter & Torres Cacoullos 2014)).

An example that involves reordering of lexical elements is the English verb-particle construction, where a transitive verb-particle lexical construction may occur as two adjacent elements preceding the object NP, or may embrace the object NP:

(15) *English*

pick	up	[the clothes]	~ pick	[the clothes]	up
V	Part	NP			

(Haddican et al. 2020; Röthlisberger & Tagliamonte 2020))

There are also many examples involving 'basic word order' (SVO, SOV etc), in languages where this is relatively flexible (Dryer 2013).

Studies of variable order have revealed a range of conditioning factors such as semantics, phonology and information structure. Spanish object clitic placement (as in 15) is primarily influenced by object topicality and animacy, and the degree of grammaticalisation of the finite verb (Schwenter & Torres Cacoullos 2014: 524). SOV constituent order variation is also known to be strongly influenced by information structure (Payne 1992). English particle placement (as in 16) is primarily influenced by the phonological weight of the object

[3] In fact, these variations may not be purely a matter of sequential order, as prosodic constituency might also vary in some instances (Himmelmann 2022); however this is not usually discussed in grammars so we here focus on sequential order.

NP (Haddican et al. 2020; Röthlisberger & Tagliamonte 2020). In Tagalog, variable ordering of N and Adj in an AdjP has been shown to be strongly influenced by phonotactics at word boundaries (Shih & Zuraw 2017). In all these instances, variant selection is largely predicted by factors relating to production planning and the referential structure of discourse (Tamminga et al. 2016), but social-indexicality appears to be largely absent.

There are relatively few instances in the sociolinguistics literature where order variation is reported to differentiate dialects; but there are some. For example, although English particle verb order is primarily driven by phonology, the centuries of separation between American and British Englishes have facilitated a divergence of frequencies. Both American and British Englishes are slowly increasing their frequency of VOP, to the detriment of VPO, but this change is slightly more advanced in Britain (Haddican et al. 2020; Röthlisberger & Tagliamonte 2020). This is a kind of slow-moving, 'stochastic' dialectal divergence, which appears to be facilitated by reduced social contact, rather than being driven by group interaction and social-indexicality. On the other hand, stochastic divergence may eventually become categorical, and this may then lead to a more sociolinguistically salient variable. For example, some north-western British English dialects developed a double-object dative construction that uses a different order from other dialects (16) (Gast 2007; Siewierska & Hollmann 2007; Gerwin 2013). Because the north-western form is not used at all in other dialects, this difference may be more salient to language users, compared to a stochastic divergence.

(16) *British English*

gave	it	him ~	gave him it	(e.g. Manchester English ~ other dialects)
V	Th	Rec		

In our database there are fewer order variables (N=154) compared to form variables (N=664), though there was at least one order variable in each of the 42 grammars. A minority of these order variables (21%, N=32) are reported to differentiate dialects, though as we will see below, a clearer pattern is revealed once we break this down according to degrees of dialect contact.

Our order variables range across diverse phrase and word structures. There are several examples of variable orderings in basic word order (17), and also word order within the NP (18, 19):

(17) *!Xun transitive clause*

SVO ~ OVS	(no dialectal association; applies only to certain verbs) (Heine & König 2015: 228)

(18) *Domari quantification*

N Quant ~ Quant N	(no dialectal association) (Matras 2012: 208–209)

(19) *Kakataibo determiners*

Det N ~ N Det	(no dialectal association) (Zariquiey 2018: 44)

Other examples involve adverbs (20), clause-level functions such as negation and TAM (21, 22), and relative clauses (23). None of these examples is reported to have a dialectal association (or indeed any other form of social-indexicality).

(20) *Skolt Saami adverbial phrase*

Clause ADVP ~ ADVP Clause (no dialectal association) (Feist 2015: 282)

(21) *Hoava future negation*

NEG FUT ~ FUT NEG (no dialectal association) (Davis 2003: 243)

(22) *Madurese transitive hortative*

V O HORT ~ HORT V O (no dialectal association) (Davies 2010: 166)

(23) *Mongsen Ao relative clause*

REL NP ~ NP REL (no dialectal association) (Coupe 2007: 202, 222)

There are also some order variables (N=36) that involve the positioning of affixes and clitics. Most of these involve variation in the relative order of multiple affixes (24, 25). But there are also some that involve an affix or clitic that may attach at variable positions in the phrase (26).[4]

(24) *Bantawa dual suffix*

a. V_1-PST-DU-V_2-PST

b. V_1-PST-V_2-PST-DU (no dialectal association) (Doornenbal 2009: 274)

(25) *Urarina plural habitual*

V-PL-HAB ~ V-HAB-PL (no dialectal association) (Olawsky 2006: 523,524)

(26) *Tundra Nenets pronominal affix in relative clause*

V-Pron N ~ V N-Pron (no dialectal association) (Nikolaeva 2014: 323,329)

As noted above, a minority of order variables are dialectal. The dialectal instances are found across word, affix and clitic constituent levels. For example in Ma'di, past transitive clauses are SOV in one dialect area and SVO in another (27).[5] In Turung, speakers in some villages sometimes use a different basic word order from those in others (28). This is an instance where one dialect has a fixed order (SOV) and the other shows variation SOV~SVO, which is reported to have arisen from contact in the latter villages with a neighbouring SVO language (Morey 2010: 513). Like the English double-object dative example above (16), the

[4] Some linguists might take variable sites of attachment as evidence for clitic rather than affix status. But since there is no consensus on how to distinguish affixes from clitics (Spencer & Luis 2012: 220), in our coding we simply follow the authors of grammars in how they distinguish clitics vs affixes.

[5] There is a slight caveat: the Ma'di SOV ~ SVO variable is not quite purely syntagmatic in its variation, as there is also a difference in tonal verb inflection between the two variants, where Lokai SOV uses a non-past low tone on the verb, but 'Burolo SVO does not. However the ordering of phrases can be considered the primary dimension of variation and therefore we coded this as an order variable. Note also that both Lokai and 'Burolo have SVO order for uninflected verbs that encode present or future tense (p.541).

Turung and Ma'di examples show that dialects can sometimes be differentiated by ordering. But this is rare compared to dialectal form variables.

(27) *Ma'di transitive clause in past tense*

S O V ~ S V O (Lokai dialect ~ 'Burolo dialect) (Blackings & Fabb 2003: 174)

(28) *Turung transitive clause*

S V O ~ S O V (Tai villages ~ others) (Morey 2010: 513)

Closer inspection of dialectal word-order variables reveals that they are noticeably concentrated in a small number of languages. Almost half the instances (14/32) are accounted for by two languages: Basque and !Xun (29, 30). Both Basque and !Xun have some dialects that are very distant from one another, and reported to be mutually unintelligible; it is in these dialect relations that we find the bulk of dialectal word-order variables. This suggests that dialectal order variations tend to arise when the speakers in two dialect groups have minimal social contact.

(29) *Basque factive negation*

V.FACT V NEG Aux ~ V.FACT NEG Aux. V
(western dialects ~ others) (Hualde & Urbina 2003: 524–525)

(30) *!Xun serial verb construction*

S V_1 TAM V_2 ~ S TAM V_1 V_2
(western 'W2' dialects ~ others) (Heine & König 2015: 92)

Among the remaining nine instances of dialectal word-order variables, spread out over seven languages, a common factor is that contact with an unrelated language is mentioned as the source of the variation. This was already illustrated above for Turung (28), and another example is in the Jerusalem dialect of Domari, where possessor/possessum ordering is reported to have flipped due to intensive bilingualism with Arabic:

(31) *Domari possession*

POSSM POSSR ~ POSSR POSSM
(Jerusalem dialect ~ Syrian dialects) (Matras 2012: 168)

There are also a few instances of dialectal variation in affix order (N=6). Examples are illustrated here from Slave (32) and Bininj Gun-wok (33). An example of dialectal order variation in clitics is found in Somali (34).

(32) *Slave negated verb with incorporated post-position*

NEG-Obj-PostP-V ~ Obj-PostP-NEG-V (Hare & Slavey ~ Bearlake)
(Rice 1989: 777)

(33) *Bininj Gun-wok immediate prefix*

NPST-Subj-IMM-V ~ NPST-IMM-Subj-V
(Gun-djeihmi clan lect ~ Kunwinjku clan lect) (Evans 2003: 320)

(34) *Somali negative interrogative*

Q=Pro=NEG ~ Q=NEG=Pro (central region ~ others) (Saeed 1999: 275)

4.3. Omission variables

The third major type of grammatical variable in our data is the OMISSION variable, where the difference between two variants consists solely in the presence/absence of a grammatical marker. A well-studied omission variable is in French verbal negation, where the particle *ne* is variably present or absent (35). The single-marked version has over time become dominant in speech, and spread across geographic dialects, while the double-marked version remains in writing (Ashby 1981; Armstrong 2002; Martineau & Mougeon 2003).

(35) *French negation*

je	(ne)	sais	pas	(written ~ spoken)
1SG	NEG	know	NEG	

'I don't know'

Other well-studied examples include the presence/absence of an overt relativiser in some English relative clause types (Jaeger 2010; Wasow et al. 2011), optional case markers in Japanese (Kurumada & Jaeger 2015), and in various Australian languages (McGregor 2006; Gaby 2008; Meakins 2015). In all these instances, the omissable grammatical marker is to some extent semantically redundant, and its presence/absence is largely determined by informational context.

Omission variables are less frequent than form variables, but more frequent than order variables. They account for 23% (N=268) of all grammatical variables we identified, and at least one omission variable was identified for 40 of the 42 languages. 25% of omission variables (N=68) are dialectal, which is about the same rate as for order variables.

As in the well-studied examples above, omission variables in our data often appear to be driven by redundancy. For example in the Tundra Nenets omission variable shown earlier in this paper (3, repeated for convenience as 36), scalar comparison is expressed by the juxtaposition of two NPs, with an ablative suffix to mark the standard of comparison. Optionally, a comparative suffix may appear on the adjective denoting the scalar property, but we might assume that the comparative meaning of the construction is already clear without this marker.

(36) *Tundra Nenets comparison*

t'uku°	*pəni°*	*taki°*	*pəne-xəd°*	*səwa(-rka)*	(no dialectal association)
this	coat	that	coat.ABL	good(-COMP)	

'This coat is better than that one.' (Nikolaeva 2014: 174)

Tundra Nenets also provides one of the few examples of an omission variable with a dialectal association. Negative clauses always begin with a NEG particle, but speakers from the eastern region additionally use a *-q* 'connegative' suffix on the verb, which is often omitted by speakers from the western region. Note however that this is reported to be a stochastic, rather than categorical difference.

5. Testing the relationship between structural type and dialectal status

In the previous section we noted the percent of each variable type that is dialectal. However, we can understand these figures better by factoring out both structural types and dialect contact. Figure 4(a,b) illustrates grammatical variables, coloured to distinguish dialectal variables in orange and non-dialectal variables in orange, grouped by degrees of social distance. Figure 4a shows raw count data, and 4b shows proportions of dialectal vs non-dialectal. The figure shows that just over half of form variables are dialectal, and this tendency is quite consistent across degrees of distance. For order and omission variables, however, only a minority are dialectal in settings of Close or Medium contact, while around half are dialectal in settings of Distant contact.

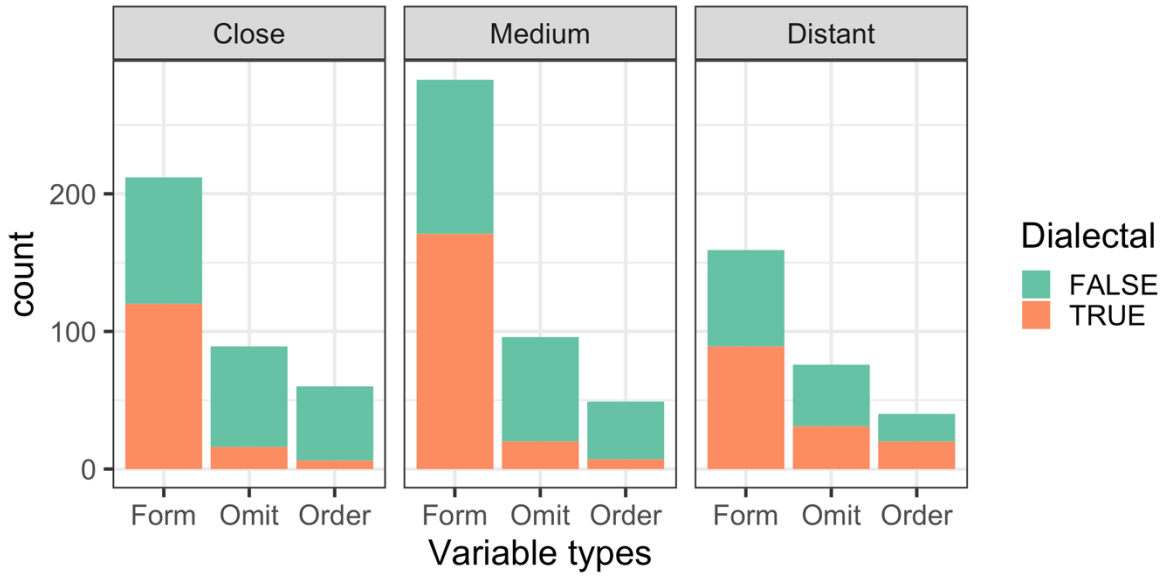

Figure 4a. Counts of grammatical variables categorised as dialectal or non-dialectal, grouped by degrees of dialect contact.

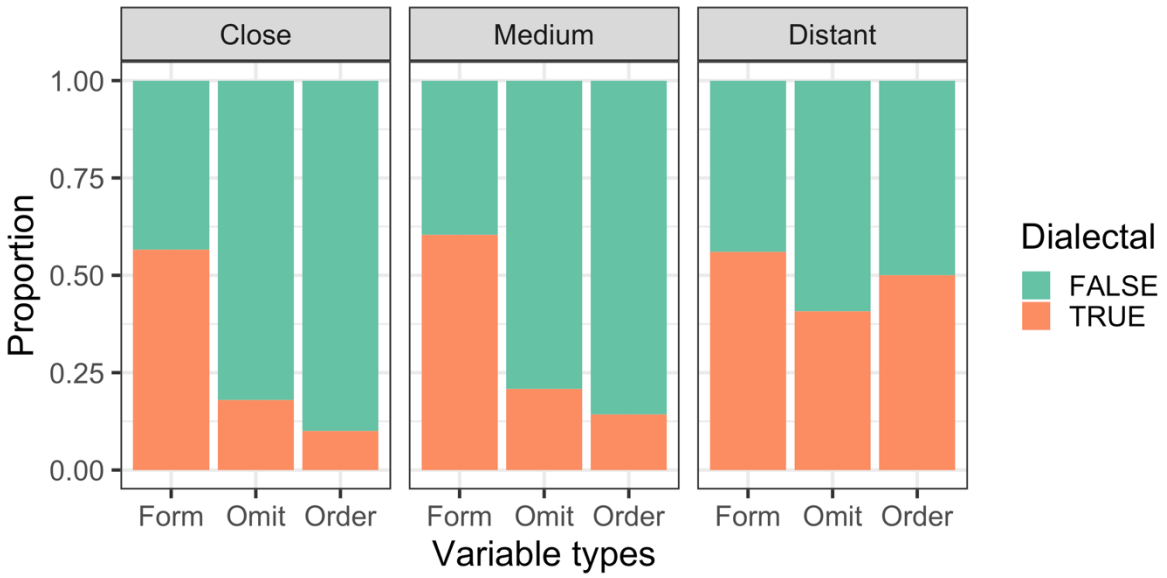

Figure 4b. Proportions of dialectal vs non-dialectal grammatical variables.

To test for a relationship between structural types, social distance and dialectal differentiation, we fitted a mixed-effects regression model as follows.[6] The outcome to be predicted is whether a grammatical variable is dialectal or non-dialectal, and the fixed effects are structural type and social distance. We modelled the fixed effects using treatment coding, with form variables, in situations of close contact, as the baseline or 'reference levels'. The model then estimates the effect of shifting to either an order or omission variable, the effect of increasing social distance, and finally an interaction effect of changing both the structural type and increasing distance. Order and omission are coded as treatment contrasts, each being compared against form variables. Distance is coded as a polynomial contrast, modelling a linear change in dialectal status as distance increases: Close < Medium < Distant (Schad et al. 2020).

Based on our impressionistic analysis of the data, we expect that omission or order types, in situations of close social contact, should have a lower probability of being dialectal. On the other hand, increasing social distance, while focusing on form variables, does not appear to affect the probability of a variable being dialectal. Finally, we expect an interaction between both omission and order variables and social contact: when we increase social distance, omission and order variables should be more likely to be dialectal, compared to their low probability of being dialectal under close contact. We also include random effects in the model to control for undue influence from particular language families.[7] As noted above, our data contains different quantities of data from different language families, but we can control for this imbalance by including a random intercept for each family, and a random slope parameter for structural type in each family.

The model was fitted in R using the lmer package (Bates et al. 2015), with estimates of the predictors shown in Table 4. The intercept represents the probability of a form variable, in a close-contact situation, being dialectal. This is not significantly different from zero, i.e. even chances of being dialectal or not. The fixed effects conform to the expectations we derived from our impressionistic analysis. Comparing omission or order types to the form baseline produces highly significant, negative effects on the probability of a grammatical expression being dialectal. The effect of social distance, when considered with the form type as a reference level, is not significant.[8] But when considering the interaction of omission or order variables with social distance, we find that greater distance increases the probability of a variable being dialectal. This is a relatively weak effect for omission variables, but rather stronger for order variables.

Table 4. Model coefficients of a mixed-effects regression predicting dialectal status

PARAMETER	ESTIMATE AND 95% CI (IN LOG ODDS)	P-VALUE SIGNIFICANCE

[6] R code and raw data used for this regression are available at [ANONYMISED REPOSITORY].
[7] Modelling individual languages produces very similar results, as most of our families are represented by a single language.
[8] lmer provides both linear and quadratic estimates of a three-level polynominal contrast. We here report whichever is the strongest estimate for each social distance factor.

Intercept (Form variables in close contact)	-0.04 [-0.57, 0.45]	-
Type: Order	-1.84 [-2.80, -1.22]	***
Type: Omission	-1.96 [-2.63, -1.41]	***
Social distance (linear)	-0.22 [-0.72, 0.27]	-
Social distance x Order (linear)	1.38 [0.47, 2.40]	**
Social distance x Omission (quadratic)	0.71 [-0.03, 1.52]	.

The family-level intercepts for language families range from -2.01 (Baining) to 2.10 (Athapaskan), with a standard deviation of 1.15. These figures represent the probability of grammatical variables being dialectal in different families (at the reference level, form variables in close contact) – for example, that most variables are reported to be dialectal in Athapaskan, and only a few in Baining. As noted above, we expect that there may be differences between grammar writers in how much attention they pay to dialectology, and we suspect that these random intercepts are more likely explained by grammar-writing methodologies than by actual differences between language families. The family-level slopes for the order type (versus form) range from -2.23 (Austronesian) to -1.41 (Athapaskan), with a standard deviation of 0.43. The family-level slopes for the omission type (again versus form) range from -2.46 (Niger-Congo) to -1.33 (Kxa), with a standard deviation of 0.46. Notice that these family level slopes are all negative, and suggesting that order and omission variables are less likely to be dialectal irrespective of language family. The difference between structural types thus appears to be a robust cross-linguistic pattern, rather than being unduly influenced by exceptional families in our data.

6. Summary of findings and implications

One simple finding of our study is that grammatical variables often differentiate dialects. When reference grammars report two distinct ways of expressing a grammatical meaning, roughly half of these are also reported to distinguish dialects, either categorically or stochastically. Although reference grammars cannot be read as comprehensive sources on dialectology, this finding nonetheless suggests that dialect differentiation is frequently facilitated by grammatical variables.

Our more important finding is on structural types of variation, and degrees of social contact. Grammatical form variables frequently differentiate dialects, and this applies equally to dialects with close or distant social contact. If we assume that dialects usually reduce their degree of contact over time, this would imply that much of the grammatical form differentiation is established during periods of close contact, and little is added once they move apart. By contrast, order and omission variables are more likely to cross-cut dialects when they are in close contact, but become more likely to differentiate dialects as they lose contact. Together these findings suggest that grammatical form variables are driven to a greater extent by social-indexicality, compared to order and omission variables that have little social-indexical function.

In the introduction we defined 'linguistic divergence' as diversification driven by contact, as opposed to drift. The implication of our findings is that linguistic divergence affects not just lexical items, but also grammatical markers such as affixes and function words. Figure 5 extends the schema from Figure 1 above, representing what we conjecture to be typical pathways for form and order variables in linguistic divergence. An initially integrated social group splits into two, and the degree of social contact between these groups (dotted lines) gradually decreases over time. Form variables tend to differentiate dialects soon after group fission, while there is still regular social interaction between members of the groups. The grammatical form difference persists even after social contact wanes, as a relic of the earlier phase. By contrast, order variables only begin to differentiate groups once social contact wanes.

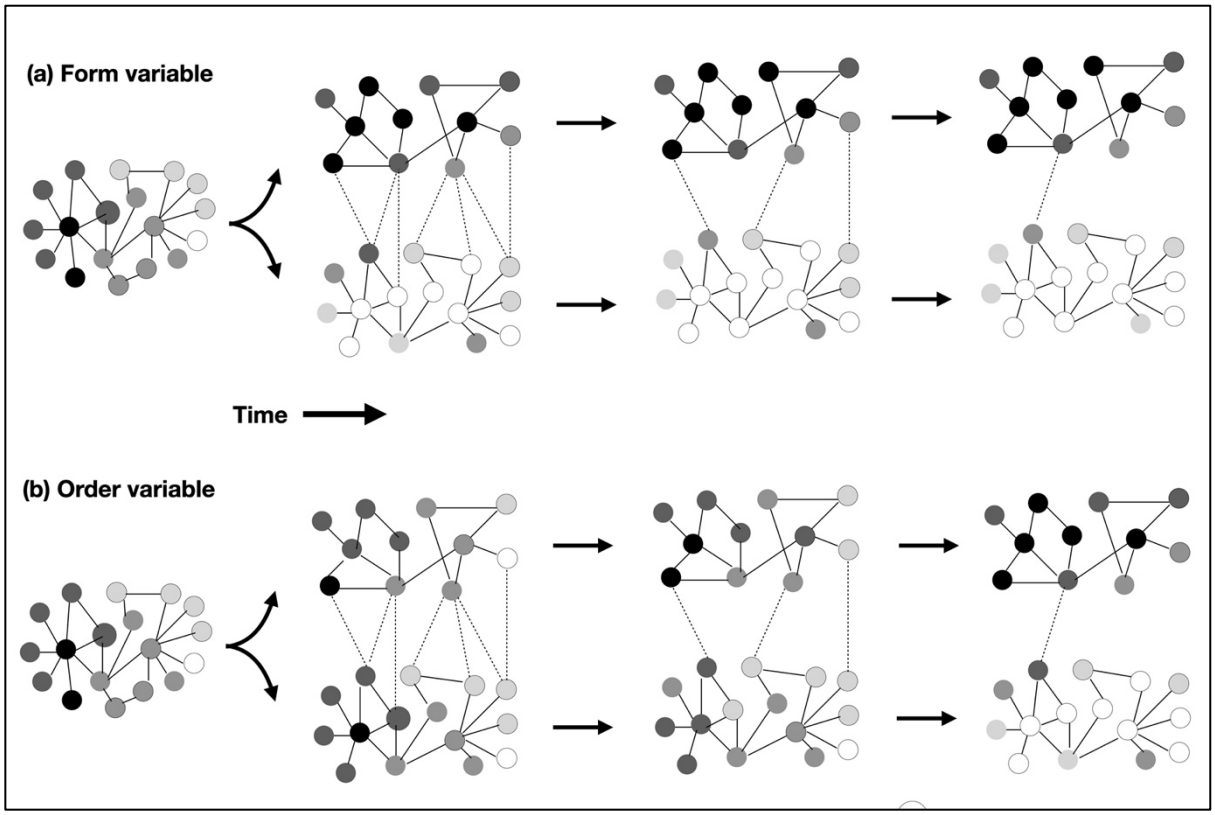

Figure 5. Schematic model of how form and order variables differentiate dialects over time.

If we project this schema onto millenia-old language families, it would suggest that families diversifying 'in-situ', with prolonged social contact between lects, should exhibit more diversity in the forms of grammatical markers compared to families that spread in a more dispersed manner. We hope that future research will be able to test this conjecture, for example by enriching phylogenetic data with information on (historical) degrees of social contact. One preliminary study of this type has investigated contact and lexical divergence in Oceanic languages (Miceli et al. 2016), finding some evidence that more social contact favours more diversification of basic vocabulary.

If social signalling is more easily achieved with grammatical markers than with linear ordering, this would quite consistent with sociolinguistic research. There are many well-

studied examples of form variables that are salient markers of social identity, such as *isn't~ain't* in English, or *voseo* in South American Spanish. For order variables, on the other hand, it is more difficult to point at sociolinguistically salient examples, though there are some rare cases such as British English *give me it ~ give it me*.

6.1. Towards a deeper explanation

Why exactly should order variation be less likely to develop socio-indexicality, compared to form variation? In the discussion above it was noted that order variables, where they have been studied in detail, have been shown to be strongly influenced by phonological, semantic and pragmatic factors in their contexts of occurrence (Milroy & Gordon 2003: 187; Cheshire et al. 2005). One possible explanation for their lack of social-indexicality is that these strong language-internal factors inhibit the development of social signalling. If variant selection is strongly predicted by the linguistic contextual factors in each instance of occurrence, this may mean that there is less variance available for socio-indexicality.

As an example of a linguistic variable that is strongly predicted by linguistic contextual factors, take the English dative alternation (Bresnan & Nikitina 2009; Bresnan & Ford 2010). Regression modelling of speakers' choice between two alternative dative expressions shows that variant selection is influenced by linguistic factors including definiteness, discourse accessibility, animacy, identity of verb lexeme and the number of words in each constituent. The model combining these predictors achieves 94.5% accuracy on unseen corpus data (Bresnan & Ford 2010: 180), suggesting that although both variants are grammatically acceptable, there is in fact very little variance in their occurrence, once linguistic contextual factors are taken into account. The dative alternation is not known to have any type of social-indexicality, and this may be precisely because there is so little variance left over after linguistic context is factored out. Similar arguments may apply to omission variables, which are also reported to be highly conditioned by linguistic contextual factors, especially informational redundancy (Wasow et al. 2011; Kurumada & Jaeger 2015).

The argument outlined above is similar to a theory of social-indexicality in terms of expectations and surprisal (Rácz 2013; Jaeger & Weatherholtz 2016; Lai et al. 2020). Originally developed with respect to phonetic variables, the core proposal is that listeners learn the contextual probabilities of hearing various sounds. For a phonetic variant to be social-indexical, it should have a high surprisal (negative log probability) based on purely linguistic context. For example, in British English, glottalisation of stops has low surprisal in coda position – it occurs quite frequently as a function of language-internal articulatory patterns, and it therefore has little potential to be interpreted as a social-indexical marker. But in intervocalic position it has higher surprisal, facilitating social-indexicality of intervocalic t-glottalisation (Rácz 2013: 145). A theory of social-indexicality in terms of surprisal is compatible with the idea that order variables resist social-indexicality because they are so heavily conditioned by linguistic context. Again, this would imply that variant selection leaves little residual surprisal, and such surprisal is key to the interpretation of as social signalling.

References

The complete list of reference grammar sources can be found in the Supplementary Information.

Abraham, Werner. 2006. Dialect and typology: Where they meet - and where they don't. In Nevalainen, Terttu & Klemola, Juhani & Laitinen, Mikko (eds.), *Types of variation: Diachronic, dialectal and typological interfaces*, 3–17. Amsterdam: John Benjamins.

Agha, Asif. 2003. The social life of cultural value. *Language & Communication* (Words and Beyond: Linguistic and Semiotic Studies of Sociocultural Order) 23(3). 231–273. (doi:10.1016/S0271-5309(03)00012-0)

Armstrong, Nigel. 2002. Variable deletion of French ne: a cross-stylistic perspective. *Language Sciences* 24(2). 153–173. (doi:10.1016/S0388-0001(01)00015-8)

Ashby, William J. 1981. The Loss of the Negative Particle ne in French: A Syntactic Change in Progress. *Language* 57(3). 674–687. (doi:10.2307/414345)

Atkinson, Quentin D. & Meade, Andrew & Venditti, Chris & Greenhill, Simon J. & Pagel, Mark. 2008. Languages Evolve in Punctuational Bursts. *Science*. American Association for the Advancement of Science. (doi:10.1126/science.1149683)

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

Blackings, Mairi & Fabb, Nigel. 2003. *A Grammar of Ma'di*. Berlin: Mouton de Gruyter.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.

Bresnan, Joan & Ford, Marilyn. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

Bresnan, Joan & Nikitina, Tatiana. 2009. The gradience of the dative alternation. In Uyechi, Linda & Hee Wee, Lian (eds.), *Reality exploration and discovery: Pattern interaction in language and life*, 161–184. Stanford: CSLI Publications.

Broadwell, George Aaron. 2006. *A Choctaw reference grammar*. Lincoln, NE: University of Nebraska Press.

Campbell-Kibler, Kathryn. 2010. The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*. Cambridge University Press 22(3). 423–441. (doi:10.1017/S0954394510000177)

Chambers, Jack & Trudgill, Peter. 1998. *Dialectology*. Second edition. Cambridge: Cambridge University Press.

Cheshire, Jenny. 1987. Syntactic variation, the linguistic variable, and sociolinguistic theory. *Linguistics* 25(2). 257–282. (doi:10.1515/ling.1987.25.2.257)

Cheshire, Jenny & Kerswill, Paul & Williams, Ann. 2005. Phonology, grammar, and discourse in dialect convergence. In Auer, Peter & Hinskens, Frans & Kerswill, Paul (eds.), *Dialect Change: Convergence and Divergence in European Languages*, 135–168. Cambridge: Cambridge University Press.

Collazos, Ana María Díaz. 2015. *Desarrollo sociolingüístico del voseo en la región andina de Colombia (1555–1976)*. De Gruyter. (https://www.degruyter.com/document/doi/10.1515/9783110404142/html) (Accessed June 8, 2021.)

Coupe, Alexandre R. 2007. *A grammar of Mongsen Ao*. Berlin: De Gruyter.

Croft, William. 2000. *Explaining language change: An evolutionary approach*.

Davies, William D. 2010. *A grammar of Madurese*. Berlin: Mouton de Gruyter.

Davis, Karen. 2003. *A grammar of the Hoava language, Western Solomons*. Canberra: Pacific Linguistics.

Dediu, Dan & Cysouw, Michael & Levinson, Stephen C. & Baronchelli, Andrea & Christiansen, Morten H. & Croft, William & Evans, Nicholas et al. 2013. Cultural evolution of language. In Richerson, Peter J. & Christiansen, Morten H. (eds.), *Cultural evolution: Society, technology, language, and religion*, 303–332. Cambridge, MA: MIT Press.

Di Carlo, Pierpaolo. 2018. Towards an understanding of African endogenous multilingualism: ethnography, language ideologies, and the supernatural. *International Journal of the Sociology of Language*. De Gruyter Mouton 2018(254). 139–163. (doi:10.1515/ijsl-2018-0037)

Di Garbo, Francesca & Kashima, Eri & Napoleão de Souza, Ricardo & Sinnemäki, Kaius. 2021. Concepts and methods for integrating language typology and sociolinguistics. In Ballarè, Silvia & Inglese, Guglielmo (eds.), *Tipologia e Sociolinguistica* (Nuova Serie), 143–176. Milano: Officinaventuno. (doi:10.17469/O2105SLI000005)

Dixon, R.M.W. 1988. *A grammar of Boumaa Fijian*. Chicago: University of Chicago Press.

Döhler, Christian. 2018. *A grammar of Komnzo*. Berlin: Language Science Press. (doi:10.5281/zenodo.1477799) (http://langsci-press.org/catalog/book/212) (Accessed April 19, 2019.)

Doornenbal, Marius. 2009. *A grammar of Bantawa*. Meteren: Netherlands Graduate School of Linguistics.

Dryer, Matthew S. 2013. Order of subject, object and verb. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://wals.info/chapter/81)
Dryer, Matthew S. & Haspelmath, Martin (eds.). 2005. *The world atlas of language structures*. Oxford: Oxford University Press.
Dunbar, Robin I. M. 2003. The Origin and Subsequent Evolution of Language. *Language Evolution*, 219–234. Oxford: Oxford University Press. (doi:10.1093/acprof:oso/9780199244843.003.0012)
Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4). 453–476.
Eckert, Penelope. 2019. The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language* 95(4). 751–776. (doi:10.1353/lan.2019.0072)
Ellison, T. Mark & Miceli, Luisa. 2017. Language monitoring in bilinguals as a mechanism for rapid lexical divergence. *Language* 93(2). 255–287. (doi:10.1353/lan.2017.0014)
Epps, Patience. 2008. *A grammar of Hup*. Berlin: Mouton de Gruyter.
Epps, Patience. 2020. Amazonian linguistic diversity and its sociocultural correlates. In Crevels, Milly & Muysken, Pieter (eds.), *Language dispersal, diversification, and contact: A global perspective*. Oxford: Oxford University Press. (doi:10.1093/oso/9780198723813.003.0016)
Errington, J. Joseph. 1985. On the nature of the sociolinguistic sign: Describing the Javanese speech levels. In Mertz, Elizabeth & Parmentier, Richard J. (eds.), *Semiotic mediation: Sociocultural and psychological perspectives*, 287–310. New York: Academic Press.
Evans, Nicholas. 2003. *Bininj Gun-Wok: a pan-dialectal grammar of Mayali, Kunwinjku and Kune*. Canberra: Pacific Linguistics.
Evans, Nicholas. 2019. Linguistic divergence under contact. In Cennamo, Michela & Fabrizio, Claudio (eds.), *Historical Linguistics 2015: Selected papers from the 22nd International Conference on Historical Linguistics*, 564–591. Amsterdam: John Benjamins. (https://benjamins.com/catalog/cilt.348.26eva) (Accessed January 5, 2022.)
Feist, Timothy. 2015. *A grammar of Skolt Saami*. Helsinki: Suomalais-Ugrilainen Seura.
Fernández Acosta, Diana. 2020. El voseo en Medellín, Colombia: Un rasgo dialectal distintivo de la identidad paisa. *Dialectología* 24. 91–109.
Ford, Lysbeth. 1998. *A description of the Emmi language of the Northern Territory of Australia*. Canberra: Australian National Univeristy. (PhD thesis.) (https://openresearch-repository.anu.edu.au/handle/1885/10796)
François, Alexandre. 2014. Trees, waves and linkages: Models of language diversification. In Bowern, Claire & Evans, Bethwyn (eds.), *The Routledge handbook of historical linguistics*, 161–189. London: Routledge.
Gaby, Alice. 2008. Pragmatically case-marked: Non-syntactic functions of the Kuuk Thaayorre ergative suffix. In Mushin, Ilana & Baker, Brett (eds.), *Discourse and grammar in Australian Languages*, 111–134.
Garde, Murray. 2008. Kun-dangwok: "clan lects" and Ausbau in western Arnhem Land. *International Journal of the Sociology of Language* 191. 141–169.
Garrett, Andrew. 2006. Convergence in the formation of Indo-European subgroups: Phylogeny and chronology. In Forster, Peter & Renfrew, Colin (eds.), *Phylogenetic methods and the prehistory of languages*, 139–151. Cambridge: McDonald Institute for Archaeological Research.
Gast, Volker. 2007. I gave it him — on the motivation of the 'alternative double object construction' in varieties of British English. *Functions of Language* 14(1). 31–56.
Gerwin, Johanna. 2013. Give it me!: Pronominal ditransitives in English dialects. *English Language & Linguistics*. Cambridge University Press 17(3). 445–463. (doi:10.1017/S1360674313000117)
Grace, George W. 1981. *An essay on language*. Columbia, SC: Hornbeam Press.
Greenhill, Simon J. & Wu, Chieh-Hsi & Hua, Xia & Dunn, Michael & Levinson, Stephen C. & Gray, Russell D. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*. (doi:10.1073/pnas.1700388114) (https://www.pnas.org/content/early/2017/10/02/1700388114) (Accessed October 5, 2020.)
Gumperz, John J & Wilson, Robert. 1971. Convergence and creolization: a case from the Indo-Aryan/Dravidian border in India. In Hymes, Dell (ed.), *Pidginization and creolization of languages*, 151–167. Cambridge University Press.
Haddican, Bill & Johnson, Daniel Ezra & Wallenberg, Joel & Holmberg, Anders. 2020. Variation and change in the particle verb alternation across English dialects. In Beaman, Karen V. & Buchstaller, Isabelle & Fox, Sue & Walker, James A. (eds.), *Advancing socio-grammatical variation and change*, 15–31. London: Routledge. (doi:10.4324/9780429282720-4) (https://www.taylorfrancis.com/https://www.taylorfrancis.com/chapters/edit/10.4324/9780429282720-4/towards-integrated-model-perception-erez-levon-isabelle-buchstaller-adam-mearns) (Accessed April 24, 2021.)

Haugen, Einar. 1988. Dialects as stepping stones to a language. In Thomas, Alan R. (ed.), *Methods in dialectology*, 666–673. Clevedon: Multilingual Matters.
Heine, Bernd & König, Christa. 2015. *The !Xun language: A dialect grammar of Northern Khoisan*. Köln: Rüdiger Köppe Verlag.
Hellwig, Birgit. 2019. *A grammar of Qaqet*. Berlin: Mouton de Gruyter.
Himmelmann, Nikolaus P. 2022. Prosodic phrasing and the emergence of phrase structure. *Linguistics*. De Gruyter Mouton. (doi:10.1515/ling-2020-0135) (https://www.degruyter.com/document/doi/10.1515/ling-2020-0135/html) (Accessed February 21, 2022.)
Hinskens, Frans. 1998. Variation studies in dialectology and three types of sound change. *Sociolinguistica*. Walter de Gruyter 12(1998). 155–193. (doi:10.1515/9783110245172.155)
Hualde, José Ignacio & Urbina, Jon Ortiz de. 2003. *A grammar of Basque*. Berlin: Mouton de Gruyter.
Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology* 61(1). 23–62. (doi:10.1016/j.cogpsych.2010.02.002)
Jaeger, T. Florian & Weatherholtz, Kodi. 2016. What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology* 7. (doi:10.3389/fpsyg.2016.01115) (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4971435/) (Accessed May 28, 2019.)
Kerswill, Paul & Williams, Ann. 2002. "Salience" as an explanatory factor in language change: Evidence from dialect levelling in urban England. In Jones, Mari C. & Esch, Edith (eds.), *Language change: The interplay of internal, external and extra-linguistic factors*, 81–110. Berlin: Mouton de Gruyter.
Kurumada, Chigusa & Jaeger, T. Florian. 2015. Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83. 152–178. (doi:10.1016/j.jml.2015.03.003)
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Labov, William. 1993. The unobservability of structure and its linguistic consequences. University of Ottowa.
Lai, Wei & Rácz, Péter & Roberts, Gareth. 2020. Experience with a linguistic variant affects the acquisition of its sociolinguistic meaning: An alien-language-learning experiment. *Cognitive Science* 44(4). e12832. (doi:10.1111/cogs.12832)
Lavandera, Beatriz R. 1978. Where does the sociolinguistic variable stop? *Language in Society*. Cambridge University Press 7(2). 171–182. (doi:10.1017/S0047404500005510)
Levinson, Stephen C. 1988. Conceptual problems in the study of regional and cultural style. In Dittmar, N. & Schlobinski, P. (eds.), *The sociolinguistics of urban vernaculars*, 161–190. Berlin: Mouton de Gruyter.
Levon, Erez & Buchstaller, Isabelle. 2015. Perception, cognition, and linguistic structure: The effect of linguistic modularity and cognitive style on sociolinguistic processing. *Language Variation and Change* 27(3). 319–348. (doi:10.1017/S0954394515000149)
Liddicoat, Anthony. 1994. *A grammar of the Norman French of the Channel Islands: The dialects of Jersey and Sark*. Berlin: Mouton de Gruyter.
Mansfield, John Basil & Stanford, James N. 2017. Documenting sociolinguistic variation in lesser-studied indigenous communities: Practical methods and solutions. In Hildebrandt, Kristine A. & Jany, Carmen & Silva, Wilson (eds.), *Documenting Variation in Endangered Languages* (Language Documentation & Conservation Special Publication 13), 116–136. Honolulu: University of Hawai'i Press.
Martineau, France & Mougeon, Raymond. 2003. A Sociolinguistic Study of the Origins of ne Deletion in European and Quebec French. *Language* 79(1). 118–152.
Matras, Yaron. 2012. *A grammar of Domari*. Berlin: De Gruyter.
Matsumae, Hiromi & Ranacher, Peter & Savage, Patrick E. & Blasi, Damián E. & Currie, Thomas E. & Koganebuchi, Kae & Nishida, Nao et al. 2021. Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Science Advances*. American Association for the Advancement of Science 7(34). eabd9223. (doi:10.1126/sciadv.abd9223)
McGregor, William. 2006. Focal and optional ergative marking in Warrwa (Kimberley, Western Australia). *Lingua* 116. 393–423.
Meakins, Felicity. 2015. From absolutely optional to only nominally ergative: The life cycle of the Gurindji ergative suffix. In Gardani, Francesco, F. & Arkadiev, Peter M. & Amiridze, Nino (eds.), *Borrowed Morphology*, 189–218. Berlin: Mouton de Gruyter.
Meyerhoff, Miriam & Walker, James A. 2012. Grammatical variation in Bequia (St Vincent and the Grenadines). *Journal of Pidgin and Creole Languages* 27. 209–234.
Miceli, Luisa & Ellison, T. Mark & Evans, Bethwyn & Greenhill, Simon J. 2016. Can we identify bilingual-led lexical differentiation in Oceanic? *Australian Linguistics Society Conference*. Melbourne: Monash University.
Milroy, Lesley & Gordon, Matthew. 2003. *Sociolinguistics: Method and Interpretation*. 2nd edition. Oxford: Blackwell.

Morey, Stephen. 2010. *Turung: A variety of Singpho language spoken in Assam*. Canberra: Pacific Linguistics.
Morphy, Frances. 1977. Language and moiety: Sociolectal variation in a Yu:lngu language of North-east Arnhem Land. *Canberra Anthropology* 1. 51–60.
Nettle, Daniel & Dunbar, Robin I. M. 1997. Social markers and the evolution of reciprocal exchange. *Current Anthropology*. The University of Chicago Press 38(1). 93–99. (doi:10.1086/204588)
Nikolaeva, Irina. 2014. *A grammar of Tundra Nenets*. *A Grammar of Tundra Nenets*. De Gruyter Mouton. (https://www.degruyter.com/document/doi/10.1515/9783110320640/html) (Accessed July 7, 2021.)
Olawsky, Knut J. 2006. *A grammar of Urarina*. Berlin: Mouton de Gruyter.
Paul, Hermann. 1888. *Principles of the history of language*. Translated from the second edition. London: Swan Sonnenschein, Lowrey & Co. (Trans. Strong, H.A.)
Payne, Doris L. (ed.). 1992. *Pragmatics of word order flexibility*. Amsterdam: John Benjamins.
Peterson, John. 2010. *A grammar of Kharia: A South Munda language*. Leiden: Brill.
Rácz, Péter. 2013. *Salience in sociolinguistics*. De Gruyter Mouton. (https://www.degruyter.com/document/doi/10.1515/9783110305395/html) (Accessed March 31, 2021.)
Rice, Keren. 1989. *A grammar of Slave*. Berlin: Mouton de Gruyter.
Roberts, Gareth. 2010. An experimental study of social selection and frequency of interaction in linguistic diversity. In Galantucci, Bruno & Garrod, Simon (eds.), *Experimental Semiotics: A new approach for studying the emergence and the evolution of human communication*, 138–159. Amsterdam: John Benjamins. (https://benjamins.com/catalog/is.11.1.06rob) (Accessed June 14, 2021.)
Romaine, Suzanne. 1981. On the problem of syntactic variation: a reply to Beatriz Lavandera and William Labov. *Sociolinguistic Working Papers* 82. 1–38.
Ross, Malcolm. 1996. Contact-induced Change and the Comparative Method: Cases from Papua New Guinea. In Durie, Mark & Ross, Malcolm (eds.), *The Comparative method reviewed*, 180–217.
Ross, Malcolm. 2001. Contact-induced change in Oceanic languages in North-west Melanesia. In Aikhenvald, Alexandra Y. & Dixon, R. M. W. (eds.), *Areal diffusion and genetic inheritance: Problems in comparative linguistics*, 134–166. Oxford: Oxford University Press.
Röthlisberger, Melanie & Tagliamonte, Sali A. 2020. The social embedding of a syntactic alternation: Variable particle placement in Ontario English. *Language Variation and Change* 32(3). 317–348. (doi:10.1017/S0954394520000174)
Saeed, John. 1999. *Somali*. Amsterdam: John Benjamins.
Saussure, Ferdinand de. 1959. *Course in general linguistics*. New York: The Philosophical Society. (Trans. Baskin, Wade.)
Schad, Daniel J. & Vasishth, Shravan & Hohenstein, Sven & Kliegl, Reinhold. 2020. How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language* 110. 104038. (doi:10.1016/j.jml.2019.104038)
Schwenter, Scott A. & Torres Cacoullos, Rena. 2014. Competing constraints on the variable placement of direct object clitics in Mexico City Spanish. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics* 27(2). 514–536. (doi:10.1075/resla.27.2.13sch)
Shih, Stephanie S. & Zuraw, Kie. 2017. Phonological conditions on variable adjective and noun word order in Tagalog. *Language: Phonological Data and Analysis* 93(4). e317–e352. (doi:10.1353/lan.2017.0075)
Siewierska, Anna & Hollmann, Willem B. 2007. Ditransitive clauses in English with special reference to Lancashire dialect. In Hannay, Mike & Steen, Gerard J. (eds.), *Structural-functional studies in English grammar: In honour of Lachlan Mackenzie*, 83–102. Amsterdam: Benjamins.
Silverstein, Michael. 1981. The limits of awareness. *Sociolinguistic Working Papers* 84.
Silverstein, Michael. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & Communication* 23. 193–229.
Smith, Ian & Johnson, Steve. 2000. Kugu Nganhcara. In Dixon, R.M.W. & Blake, Barry J. (eds.), *The handbook of Australian languages*, vol. 5, 355–490. Melbourne: Oxford University Press.
Sneller, Betsy & Roberts, Gareth. 2018. Why some behaviors spread while others don't: A laboratory simulation of dialect contact. *Cognition* 170. 298–311. (doi:10.1016/j.cognition.2017.10.014)
Spencer, Andrew & Luis, Ana R. 2012. *Clitics: An introduction*. Cambridge: Cambridge University Press.
Stanford, James N. 2009. "Eating the food of our place": Sociolinguistic loyalties in multidialectal Sui villages. *Language in Society* 38. 287–309.
Stanford, James N. 2016. A call for more diverse sources of data: Variationist approaches in non-English contexts. *Journal of Sociolinguistics* 20(4). 525–541.
Tamminga, Meredith & MacKenzie, Laurel & Embick, David. 2016. The dynamics of variation in individuals. *Linguistic Variation* 16(2). 300–336.
Trudgill, Peter. 1986. *Dialects in contact*. Oxford: Blackwell.
Vaughan, Jill. 2018. "We talk in saltwater words": Dimensionalisation of dialectal variation in multilingual Arnhem Land. *Language & Communication* 62. 119–132.

Wasow, Thomas & Jaeger, T. Florian & Orr, David. 2011. Lexical variation in relativizer frequency. In Simon, H. & Wiese, H. (eds.), *Expecting the unexpected: Exceptions in grammar*, 175–195. Berlin: De Gruyter.
Weinreich, Uriel & Labov, William & Herzog, Marvin. 1968. Empirical foundations for a theory of language change. In Lehmann, W. & Malkiel, Y. (eds.), *Directions for historical linguistics*, 95–198. Austin: University of Texas Press.
Wolfram, Walt. 1969. *A sociolinguistic description of Detroit negro speech*. Washington, D.C: Center for Applied Linguistics.
Zariquiey, Roberto. 2011. Aproximación dialectológica a la lengua cashibo-cacataibo (pano). *Lexis* XXXV(1). 5–46.
Zariquiey, Roberto. 2018. *A grammar of Kakataibo*. Berlin: Mouton de Gruyter.