

Dialect differences and linguistic divergence

A crosslinguistic survey of grammatical variation

John Mansfield | ORCID: 0000-0003-1167-1136

School of Languages and Linguistics, University of Melbourne,
Melbourne, Australia

john.mansfield@unimelb.edu.au

Henry Leslie-O'Neill | ORCID: 0000-0002-1470-0281

Australian National University, Canberra, Australia

Henry.Leslie-ONeill@anu.edu.au

Haoyi Li | ORCID: 0000-0002-7803-7511

Australian National University, Canberra, Australia

Haoyi.Li@anu.edu.au

Received 31 March 2022 | Accepted 16 March 2023 |

Published online 2 May 2023

Abstract

This article presents a new type of comparative linguistic survey, analyzing (socio)linguistic variation in a database of 1,155 grammatical constructions drawn from 42 diverse languages. We focus in particular on variation in the expression of grammatical meanings, and the extent to which grammatical variation differentiates geographic dialects. This is the first study we know of to present a systematic, crosslinguistic survey of dialect differentiation. We identify three main structural types of grammatical variation—*form*, *order*, and *omission*—and find that in situations of close contact between dialects, where signaling of distinct group identities is more relevant, 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 survey suggests that social signaling may have a substantial role in the divergence of grammars, and provides systematic support for previous proposals regarding convergence and divergence under contact.

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. Variationist sociolinguistics offers detailed studies of social identity signaling, but has thus far provided relatively limited crosslinguistic comparison (Stanford, 2016; Di Garbo et al., 2021). 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 signaling 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 language varieties to gradually become more different from one another (Paul, 1888: 25). 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 egalitarian multilingualism have shown that social groups may live in very close contact, while nonetheless carefully policing their language borders, deploying social norms and interactional etiquette to maintain or enhance the distinctness of their language varieties (e.g., François, 2011; 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 and 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 and 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). All these studies 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: grammatical meanings or functions that can be expressed in more than one way (e.g., in English, the 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 develop a simple structural typology of grammatical variation, and ask whether dialects are more likely to be differentiated by some types 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. While our method offers a substantially new type of data on language dynamics, it also has natural limitations as it is essentially a convenience sample of whichever grammatical variables the authors of reference grammars happen to mention. Nonetheless, the method provides valuable data for answering the specific question of whether certain types of variation are more likely to be dialectal than others.

Our typology identifies three structural types of grammatical variation, exemplified in (1)–(3) below: (a) *form variables*, in which distinct grammatical markers appear in the same linear position; (b) *order variables*, involving the same linguistic elements in different linear orders; and (c) *omission variables*, distinguished by the presence vs. absence of a grammatical marker.¹ Form

1 Grammatical glosses used throughout this article follow the Leipzig Glossing Rules. Abbreviations used: 1 first person; 2 second person; 3 third person; AAG anti-agentive; ABIL ability; ABL ablative; AGT agent; AN “action nominal”; ANAPH anaphora; AUG augment; AUX auxiliary; CAUSE cause; CLAN.INDEX clan index; COM comitative; COMP complementizer; DEF definite; DET determiner; DU dual; EMPH emphatic; EXCL exclusive; FUT future; HAB habitual; HORT hortative; ILL illative; IMM immediate; IMP imperative; INCL inclusive; INTNS intensifier; IRR irrealis; LIM limitative; LOC locative; M masculine; MID middle; NEG negative; NMLZ nominalizer; NPST non-past; NRL non-relational; OBJ object; OBL oblique; PASS passive; PFV perfective; PL plural; PRED predicative; PROX proximal; PRS present; PST past; Q question; REM remote past; S subject; SG singular; TNS tense; TOP topic; TR transitive.

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 (Smith and Johnson, 2000: 393)
 - a. *thuli-ra*
woomera-COM
'with a woomera'
 - b. *thuli-nta*
woomera-COM
'with a woomera'
- (2) Order variable: Komnzo adjectival attribution (Döhler, 2018: 89)
 - a. *zagr karfo*
distant village
'distant village'
 - b. *karfo zagr*
village distant
'distant village'
- (3) Omission variable: Tundra Nenets comparison (Nikolaeva, 2014: 174)
t'uku° pəni° taki° pəne-xəd° səwa(-rka)
this coat that coat-ABL good(-COMP)
'This coat is better than that one.'

Our main finding is that grammatical form variables frequently differentiate dialects, including dialects in close contact (in example (1), the variants are associated with the neighbouring dialects of Uwanh and Iyanh, respectively). By contrast, order and omission variables 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 conceptualizing dialect groups, linguistic variables, and social signaling. Section 3 introduces the database used for this study. Section 4 describes the three structural types of grammatical variation, 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 linguistic divergence. Section 6 summarizes our findings and discusses implications for further research.

2 Dialect differences, social contact, and social signaling

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 recognized as distinct varieties. We call such relations “dialectal” when they are based on geography (e.g., different villages, provinces, or regions), as opposed to other types of language variety associated with socioeconomic groups, subcultures, formal registers, and so on. Note that this approach to dialects is independent of ethnolinguistic 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, two very similar varieties associated with neighboring clan estates in northern Australia, which do not share an ethnolinguistic label (Ford, 1998).

The concept of “dialect” has often been used for unwritten regional varieties in relation to written, supra-regional “standard” varieties (see, e.g., Haugen, 1988; Chambers and 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.² 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.³

Dialectal differences develop through the identification of social groups with distinct geographic territories, provoking fission 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: 23–25; 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, 40 families from Jer-

2 Fijian is the only language included in this study for which dialect variables are reported to be in a standard vs. vernacular relationship. This involves Standard Fijian, based on the dialect of Bau but used as a lingua franca across Fiji, and Boumaa Fijian, spoken in a region on the island of Taveuni (Dixon, 1988).

3 Recent lexicostatistical research provides a very different approach to dialectal relations, distinguishing dialect pairs from different-language pairs based on degrees of phonological distinction between their vocabularies, as measured by orthographic Levenshtein distances (Wichmann, 2019).

sey 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. Language varieties may remain in close contact for hundreds or thousands of years, remaining similar because they maintain continual contact and share many linguistic innovations. There are also instances where dialects are in a process of convergence, rather than divergence (Trudgill, 1986). More generally, 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. But irrespective of these diverse histories, dialect differentiation is an important first step in the larger process of diversification that eventually leads to radically distinct languages. In this study 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 language change (see Section 6).

2.1 *Variables and dialects*

We conceptualize 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 select just one variant or the other.

We assume that individuals form geographically associated social groups, which are groups of individuals who associate to a geographic region and tend to 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 1a shows two social groups, one above and one below. Within each group there are dense social connections (solid lines), but there are also some social connections running between groups (dotted lines). Each individual is shaded in greyscale, representing their probability of using variants x_1 and x_2 (cf. Blythe and Croft, 2021). In Fig. 1a, we see that most individuals use both variants of this variable, and the two groups have similar distributions. We can say that this is an “intra-group” or “non-dialectal” variable. In Fig. 1b, we see the same pair of social 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 the variable shown in Fig. 1b is a “dialectal” variable—noting that it does not require a categorical split between groups,

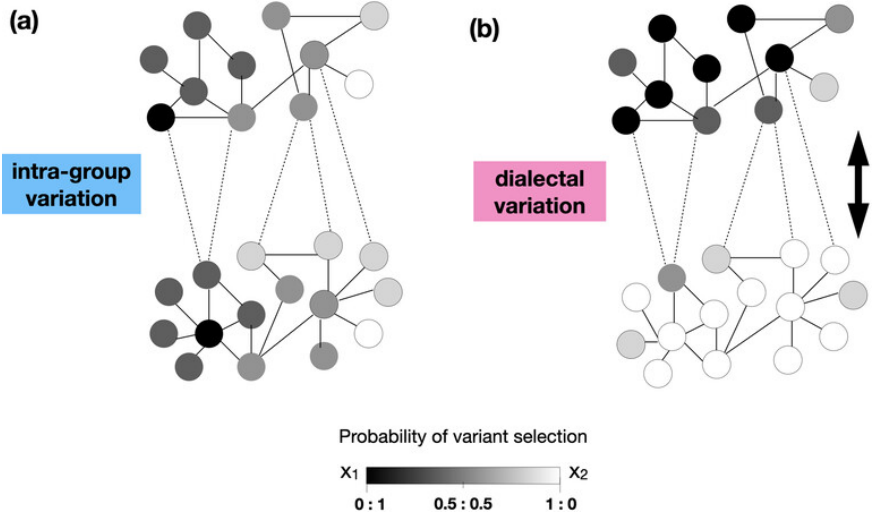


FIGURE 1 Linguistic variables and social groups. (a) Where the selection of variants $x_1 \sim x_2$ has a similar distribution in two social groups, we call this “intra-group” or “non-dialectal” variation. (b) Where the selection of variants $x_1 \sim x_2$ aligns with social groups, we call this “dialectal”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

but only that there is a notable difference between groups with respect to the variable.

2.2 Social signaling and language structure

In Fig. 1, we used dotted lines to represent interaction between individuals in different social groups. Where this interaction is substantial, it is plausible that dialect differences provide 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 subconscious 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 facilitated by social contact between groups who speak different varieties (e.g., Di Carlo, 2018; Evans, 2019; Epps, 2020). Thus linguistic differentiation may be driven by social proximity, rather than social distance (Gal, 2016: 127); and dialects, which may eventually become different enough to be counted as separate languages, are the product of “active differentiation among local communities” (François, 2012: 92). This is the linguistic instantiation of a more general anthropological process of contact-driven “schismogenesis” (Bateson, 1935). Proximity and social contact provide the context for social signaling, or “social indexicality,” whereby linguistic cues are associated with distinctive social groups (Agha, 2003; Silverstein, 2003;

Eckert, 2008; Tabouret-Keller, 2017; Eckert, 2019; for historical background, see Jahr, 2017).⁴

However, dialectal differences may also develop without any social signaling. 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 social signaling; 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 signaling. There may be cognitive or communicative constraints that make it easier to associate certain types of formal distinction with social identity. Various terms have been applied to this, such as “marker vs. indicator” (Labov, 1972), “metapragmatic awareness” (Silverstein, 1981), “pragmatic salience” (Errington, 1985), and “sociolinguistic salience” (Kerswill and Williams, 2002; Rácz, 2013; Levon and Buchstaller, 2015). Recent research has highlighted an important distinction between social indexicality and conscious awareness, which need not go hand in hand and can be difficult to disentangle (e.g., Campbell-Kibler, 2016; Drager and Kirtley, 2016). Research on sociolinguistic salience has largely focused on phonetic variation, though there is also evidence of speakers being strongly aware of at least some grammatical variables (Squires, 2016).

One enduring idea has been that “surface” linguistic forms are more capable of social signaling 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). At the same time, it can be difficult to determine when variables should be counted as surface forms, and when apparent “surface forms” actually reflect variation in underlying structure (Meyerhoff and Walker, 2012). In parallel to the socio-variationist literature, studies in comparative linguistics suggest that communities in contact tend to differentiate themselves using the forms of lexemes or grammatical markers, while subconsciously converging in their morphosyntactic structures. This again points to sociolinguistic awareness as the key factor favoring differentia-

4 There may also be purely cognitive, as opposed to sociocultural, factors at work in linguistic divergence. In a study of bilingual production, where the two languages (Dutch and English) 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 and Miceli, 2017).

tion of lexicon but not grammar (Gumperz and Wilson, 1971; Grace, 1981; Ross, 1996; Ross, 2001).

In this study we follow the lead of these earlier works in testing whether some dimensions of language have greater potential for social signaling than others. Rather than applying a surface vs. depth model (as in the socio-variationist work cited above), which depends on specific analyses of structural layers, we instead focus on paradigmatic versus syntagmatic dimensions of surface structure (see Section 4), since these can be applied in a relatively theory-neutral way based on linguistic documentation. And rather than making a lexicon vs. grammar split (as in the comparative linguistic work cited above), we focus purely on the expression of grammatical meanings, while distinguishing the structural properties of form, order, and omission (cf. François, 2011). Our findings on form vs. order variables can be roughly equated to previous findings on lexicon vs. structure under language contact, though we set aside the question of grammatico-semantic isomorphism, which has also played a major role in such studies (see Section 4.1). The most important contribution of our survey is to go beyond case studies of individual contact situations, and make generalizations about the process of linguistic differentiation based on a diverse sample. However for this, we require new methods in comparative linguistics.

3 Collating a crosslinguistic sample of dialect differences

Typological databases usually distill the information in reference grammars, in order to assign one type of grammatical expression to each language. 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 Fig. 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.⁵

Variables were added to the database by searching reference grammars for mentions of variation, using a combination of keyword searches and reading (see the supplementary materials for further details). A grammatical vari-

⁵ The database described in this section is available at <https://doi.org/10.5281/zenodo.7306236>.



FIGURE 2 Languages sampled in our database

able was coded wherever the text reports two or more ways of expressing the same grammatical meaning or function. In cases where more than two variants are documented, we annotate just two variants (see supplementary materials, Section D). Grammatical meanings are relatively abstract categories, such as future, negation, continuous aspect, first person, or directionality; or functions such as focus, subordination, or transitivity (Lehmann, 1995; Hopper and Traugott, 2003; Boye and Harder, 2012). This method yielded 1,155 grammatical variables, for each of which we coded the structural type and dialectal status into a spreadsheet, later transformed into an R data-table (see examples in Table 1 below). Most grammars also mention a range of phonological and lexical variables, which we noted for further research but have not included in this study.

Our data was selected to represent a diverse range of languages and social situations, but the sample is not formally balanced either by language family or region (see supplementary materials, Section A). Most language families are represented by a single language, but a few (mostly larger families) have multiple languages. Furthermore, the number of data points 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 (using maximal language families as annotated in Glottolog [Hammarström et al., 2022]), 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 dis-

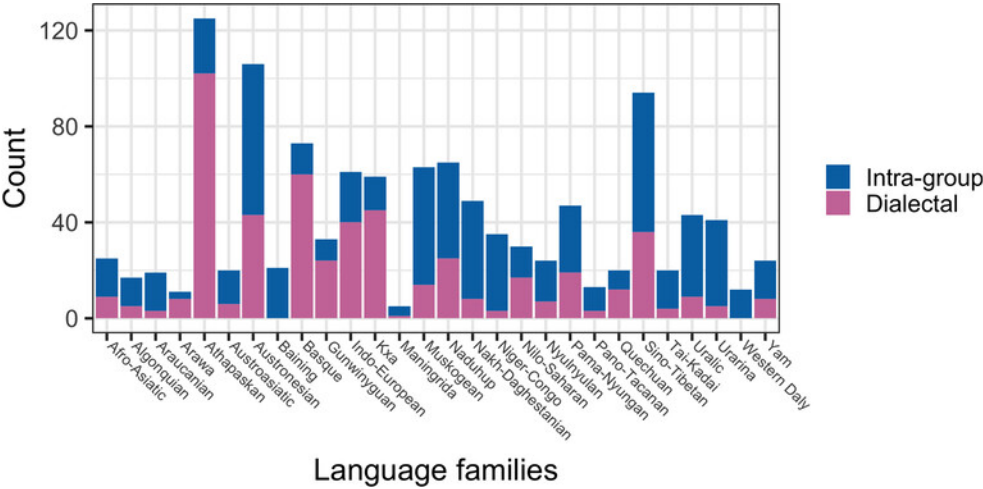


FIGURE 3 Number of grammatical variables per language family, dialectal and non-dialectal

tinct 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 and Ortiz de 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 (see Section 5). For 26 of the 28 language families, at least one of the grammatical variables was reported to be dialectal.⁶

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. It is also worth noting that some variables may involve pragmatic conditioning (especially for order variables, see Section 4.2), description of which is largely beyond the scope of the source grammars. Sociolinguists working on grammatical variation have long recognized this as a difficult problem (see, e.g., Lavandera, 1978; Romaine, 1981; Cheshire, 1987; *inter alia*).

We take an onomasiological approach, that is, using meanings as our reference points, rather than forms. Consequently a grammatical variable is anno-

6 The two exceptions are Baining (Hellwig, 2019) and Western Daly (Ford, 1998). These attest lexical and phonological differences between dialects, and report grammatical variables which are non-dialectal, but we did not identify any variables that are *both* dialectal and grammatical.

tated 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 specific future tense marker, or by a marker that spans both future and present (i.e., non-past) meanings. This is an important point to which we return below (Section 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 Language Structures (Dryer and Haspelmath, 2013). Only a subset of our grammatical variables correspond to features coded in the atlas.

Reference grammars usually attest variation in a succinct, impressionistic form, glossing over the nuances of variant distributions. In our Fig. 1b 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 categorical dialectal variables, and sometimes note that one variant is “more common” in one dialect than another. Our coding represents both of these situations as dialectal 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. Grammar writers may present something as non-dialectal or “free variation” when closer inspection would show it to be dialectal. Alternatively, grammar writers may mistakenly report something as dialectal, when it is actually intra-group variation. We must therefore assume that there is a certain degree of noise in our data sources. Furthermore, individual grammar writers have different propensities to report variation at all (as suggested by the wide range of counts in Fig. 3 above), and each has their own particular areas of interest. We therefore cannot draw any conclusions about whether certain grammatical meanings are more likely to have variable expression than others.

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, Section F). 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 absolute 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. We ask only whether certain types of grammatical variation are more likely to be dialectal than others, and the limitations of the reference-grammar method do not appear to impact on this question. The methodological limitations would invalidate our findings if there were systematic inaccuracies in whether structural types are identified as dialectal or intra-group, 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 and 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 social contact*

As well as coding multiple 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 on mutual intelligibility of dialects, geographic proximity, and so on. We used this information to create a rubric for assigning dialectal relations to three degrees of social distance: Close, Medium, and Distant (see supplementary materials, Section E). 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 parameterize 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.⁷

⁷ More precise methods of coding degrees of social contact may be a fruitful direction for future research on dialect differentiation, though it is difficult to imagine what sort of data could be practically collated for a large crosslinguistic sample. One would require quantifiable records of social intercourse, which might include marriage data, economic or migration data, or perhaps mobile phone data if it were accessible. However it would still be important to consider

The Kugu Nganhcara grammar (Smith and 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 patrilineal clans, each of which is associated with distinct geographic territory and has its own dialect or “clan lect” (Smith and 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 Islands 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 organized 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 30 km), 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 kilometers 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

the diachronic dimension, that is, not just contemporary social contact but also changes in degree of contact over time.

TABLE 1 Examples of grammatical variable coding (see supplementary materials for further details and references)

Language	Meaning	Var1	Var2	Type	Dialects	Contact
Kugu Nganhcara	Comitative	N- <i>ra</i>	N- <i>nta</i>	Form	Uwanh ~ Iyanh	Close
Kugu Nganhcara	Pronoun 1.PL.EXCL	<i>ŋaŋca</i>	<i>ŋana</i>	Form	Muminh ~ Iyanh	Close
Kugu Nganhcara	Transitive clause	SVO	SOV	Order	NA	Close
Basque	Future participle (stem ending /n, l/)	V.PTCPL- <i>ko</i>	V.PTCPL- <i>en</i>	Form	western ~ eastern	Distant
Hup	Adj-INTNS	Adj- <i>Vcap</i>	Adj- <i>icap</i>	Form	Barriera ~ Tat Deh	Close
Hup	V-INTNS	V- <i>tubud</i>	V- <i>tiud</i>	Form	others ~ western	Distant
Nishnaabemwin	Nominal conjunction	NP conj NP	NP NP	Omit	NA	Medium
Nishnaabemwin	Plural	N- <i>o:g</i>	N- <i>ag</i>	Form	NA	Medium

groups, and the central and eastern speakers say the western dialect is “hard to understand” (Epps, 2008: 13). On this basis, we assigned a Distant relationship between the western dialect and the other two.

Table 1 shows the coding of some example variables. Most of the examples shown here are dialectal variables, but Kugu Nganhcara SOV ~ SVO and Nishnaabemwin nominal conjunction and plural marking are examples of intra-group variables (and so are coded with “NA” in the Dialects column). 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.⁸ The coding of the Type column will be explained in the following sections.

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:⁹

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

8 The two Hup variables are also examples of stochastic dialectal variables: in each instance, one of the dialect groups is reported to use both variants.
9 A small residue group of variables (8% of the total) exhibit a mixture of the criteria for the three main types (see supplementary materials, Section C). These are excluded from the analysis.

- (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 a grammatical marker, 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, as in *she isn't home ~ she ain't home*, for example. This has a social signaling function, marking social class, stance, and style (Levinson, 1988; Cheshire et al., 2005). English has other well-known form variables that also have strong social connotations. 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 second person singular markers both in free pronouns and in verbal suffixes, as seen in (4). In some areas *voseo* is a salient marker of regional dialects, for example in Colombia (Díaz Collazos, 2015: 10–13; Fernández Acosta, 2020).

- (4) Colombian Spanish; e.g., Cordoba vs. Antioquia dialects (Díaz Collazos, 2015: 10)
 - a. (*tu*) *com-es*
2SG.S eat-2SG.S
'You eat.'
 - b. (*vos*) *com-és*
2SG.S eat-2SG.S
'You eat.'

A particularly flamboyant example of dialectal form variation is in Biniŋ Gunwok, 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,” used purely for social signaling.

- (5) Biniŋ Gunwok; Djordi vs. Kurulk vs. Mok clan lects (Garde, 2008: 150–154)
 - a. *yi-njarra-kinje-men*
 - b. *yi-bayid-kinje-men*

- c. *yi-buk-kinje-men*
2SG-CLAN.INDEX-cook-IMP
'You cook it!'

Form variables may emerge either from sound changes or from grammaticalization pathways. A phonologically induced example can be seen in the American English first person singular future auxiliary, where African American dialects have innovated *I'm'a*, marking a point of differentiation from other dialects, which have *I'm gonna*. 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. This criterion therefore includes some variants that are phonologically similar, such as the Kharia clitic forms in (7) below.

Form variation via grammaticalization pathways 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 US), exhibiting different grammaticalization paths in the development of the pluralizing suffix.

Form variables are the most frequent type of grammatical variable in our data, accounting for 57% (N = 654) of all variables annotated. There is at least one 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 = 380) of form variables being dialectal. Form variation of grammatical markers therefore appears to be a crosslinguistically 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.¹⁰ Examples of pronominal form variation can be found in Fijian free pronouns, as in (6), Kharia pronominal clitics on irrealis middle verbs, as in (7), and Biniñ Gun-wok verbal agreement prefixes, as in (8).

- (6) Fijian; Standard/Bau vs. Boumaa dialects (Dixon, 1988: 54)
- | | |
|----------------|---------------|
| a. <i>koya</i> | b. <i>ʔea</i> |
| 'him' | 'him' |

¹⁰ In view of the severe difficulties in distinguishing affixes, clitics, and function words (Haspelmath, 2011; Spencer and Luis, 2012), we simply follow the source materials in how specific forms are categorized.

- (7) Kharia; intra-group (Peterson, 2010: 249)
- | | |
|--------------------|-----------------------|
| a. <i>co=na=iŋ</i> | b. <i>kayom=na=iŋ</i> |
| go=MID.IRR=1SG | speak=MID.IRR=1SG |
| 'I will go.' | 'I will speak.' |
- (8) Biniŋ Gun-wok; Kunwinjku vs. Gundedjnjenghmi dialects (Evans, 2003: 20)
- | | |
|----------------------|----------------------|
| a. <i>karri-yoy</i> | b. <i>yirri-yoy</i> |
| 1.INCL.AUG-sleep.PST | 1.INCL.AUG-sleep.PST |
| 'We slept.' | 'We slept.' |

Form variation also occurs frequently in the marking of grammatical roles for nominal expressions, such as case affixes, seen in (9), and adpositions, in (10).

- (9) Basque; eastern regions vs. Bizkaian (Hualde and Ortiz de Urbina, 2003: 184)
- | | |
|---------------------|----------------------|
| a. <i>resé-kila</i> | b. <i>mutilá-gaz</i> |
| sheep-COM | boy-COM |
| 'with the sheep' | 'with the boy' |
- (10) !Xun; E1 vs. other dialects (Heine and König, 2015: 187)
- | | |
|-------------------|---------------------|
| a. <i>ts'ù sí</i> | b. <i>ts'ù íl'í</i> |
| house LOC | house LOC |
| 'in the house' | 'in/at the house' |

Form variation is also found in clause-level functions such as TAM, in (11), and in other miscellaneous forms that can be considered broadly grammatical in their meaning, such as quantification, in (12), and assent, in (13).

- (11) Choctaw imperative suffix; intra-group (Broadwell, 2006: 194)
- | | |
|----------------------|-----------------------|
| a. <i>ahpáli-cha</i> | b. <i>sa-sso-h-o:</i> |
| kiss-IMP | 1SG.II-hit-TNS-IMP |
| 'Kiss (her)' | 'Hit me!' |
- (12) Emmi; intra-group (Ford, 1998: 134)
- | | |
|-----------------|-------------------|
| a. <i>dawal</i> | b. <i>pakwutj</i> |
| 'many' | 'many' |

(13) Urarina assent; Asna vs. other dialects (Olawsky, 2006: 882)

a. *ajara*
'yes'

b. *ěehe*
'yes'

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 different methodology would be required to investigate whether dialectal differentiation is more or less likely in specific 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 FUTURE marker vs. a more general NON-PAST marker). When markers with different functional ranges distinguish dialects, this implies that the dialects are not isomorphic in their grammatico-semantic structure. Therefore, our findings on dialects being differentiated by the form of grammatical markers should *not* be interpreted as showing that dialects differ only on their "surface" forms, since the grammatico-semantic structure is also different in many instances (cf. Grace, 1981). Investigation of such non-isomorphisms may reveal further 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.¹¹ Order variables may involve the positioning of a grammatical marker, or the reordering 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; intra-group (Schwenter and Torres Cacoullos, 2014)

a. *no puede manejarlos*
NEG can.3SG.PRS manage=3PL.M.OBJ
'She can't manage them.'

b. *no los puede manejar*

¹¹ 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 here we focus on sequential order.

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; intra-group (Haddican et al., 2020; Röthlisberger and Tagliamonte, 2020)
- | | |
|--|--|
| <p>a. <i>pick up</i> [<i>the clothes</i>]</p> <p style="margin-left: 2em;">V participle NP</p> | <p>b. <i>pick</i> [<i>the clothes</i>] <i>up</i></p> |
|--|--|

Studies of variable order have revealed a range of conditioning factors such as semantics, phonology, and information structure. Spanish object clitic placement, as in (14), is primarily influenced by object topicality and animacy, and the degree of grammaticalization of the finite verb (Schwenter and Torres Cacoullós, 2014: 524). Basic constituent order (SOV, OVS, etc.) is variable in many languages, where it is generally influenced by information structure (Payne, 1992). English particle placement, as in (15), is primarily influenced by the phonological weight of the object NP (Haddican et al., 2020; Röthlisberger and Tagliamonte, 2020). In Tagalog, variable ordering of nouns with adjective modifiers has been shown to be strongly influenced by phonotactics at word boundaries (Shih and Zuraw, 2017). In all these instances, variant selection is largely determined by factors relating to production planning and the referential structure of discourse (Tamminga et al., 2016), but social signaling 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 the frequency of verb-object-particle, to the detriment of verb-particle-object, but this change is slightly more advanced in Britain (Haddican et al., 2020; Röthlisberger and Tagliamonte, 2020). This is a kind of slow-moving, stochastic dialectal drift, which appears to be facilitated by reduced social contact, rather than being driven by group interaction and social signaling. On the other hand, stochastic divergence may eventually become categorical, and this may then lead to a more sociolinguistically salient variable. For example, Dutch has variable order of a sentence-final participle and auxiliary verb, which in some areas is associated with regional dialects (De Sutter, 2005). In English, some northwestern British dialects developed a double-object dative construction that uses a different order from other dialects (Gast,

2007; Siewierska and Hollmann, 2007; Gerwin, 2013), as seen in (16). Note that this is distinct from the more familiar “dative alternation,” which involves an additional preposition.¹² Because the northwestern form in (16a) 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; e.g., Manchester vs. other dialects (Gast, 2007)

- | | | |
|-------------------|-----------|----------------------|
| a. <i>give it</i> | <i>me</i> | b. <i>give me it</i> |
| V | Theme | Recipient |

Drawing on language contact literature, we can conceptualize order variables as involving differences of *pattern*, as opposed to form variables that involve differences of *matter* (Matras and Sakel, 2007; Gardani, 2020). The contact literature investigates several types of pattern borrowing, with the general finding that contact-driven convergence affects patterns more than matter, due to social constraints that tie particular matter to a particular language (Gumperz and Wilson, 1971; Matras and Sakel, 2007: 857). Conversely, we might expect that in situations of socially mediated linguistic divergence, matter will provide more social signaling than patterns. This would therefore predict that form variables will be more likely to differentiate dialects than order variables.

In our database there are fewer order variables (N = 149) compared to form variables (N = 654), though there was at least one order variable in each of the 42 grammars. A minority of these order variables (22%, N = 33) are reported to differentiate dialects, though as we will see in Section 5, a clearer pattern is revealed once we break this down according to degrees of social contact.

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

12 The more standard English dative alternation, for example *give me it* versus *give it to me*, is not an order variable by our criteria, since one variant has an extra preposition that the other lacks (see supplementary materials, Section C). The standard alternation is well known for its linguistic conditioning factors, though one recent study also finds an association with speaker gender (Jenset et al., 2018).

- (17) !Xun basic constituents; intra-group (Heine and König, 2015: 228)

a. <i>mā</i>	<i>kōrā</i>	<i>gllú</i>	b. <i>gllú</i>	<i>kōrā</i>	<i>mí</i> ¹³
S	V	O	O	V	S
1SG exist.NEG water					
'I have no water'					

- (18) Domari quantification; intra-group (Matras, 2012: 208–209)

a. <i>šinak</i>	<i>pl-ēni</i>	b. <i>pl-ēni</i>	<i>šinak</i>
little	money-PRED.PL		
'a little money'			

- (19) Kakataibo determiners; intra-group (Zariquiey, 2018: 44)

a. <i>tsatsa</i>	<i>ěně</i>	b. <i>a</i>	<i>uni</i>
fish.sp	this	that	person
'this fish'		'that person'	

Other examples involve clause-level particles such as those for negation and TAM, seen in (20) and (21), adverbial clauses in relation to their matrix clause, in (22), or relative clauses in relation to a noun phrase, in (23). None of these examples is reported to have a dialectal association (or indeed any other form of social signaling).

- (20) Hoava; intra-group (Davis, 2003: 243)

a. *koni kipu ta-va-mate*
 FUT NEG PASS-CAUSE-die
 ‘(He) will not be killed.’

b. *kipu koni tavetí ria ba sara*
 NEG FUT make.TR.3PL 3PL EMPH those
 ‘They will not make *those*.’

- (21) Madurese; intra-group (Davies, 2010: 166)

a. <i>mara</i>	<i>baca</i>	<i>buku</i>	<i>reya!</i>	b. <i>buku</i>	<i>reya</i>	<i>baca</i>	<i>mara!</i>
HORT	read	book	this				
'Let's read this book!'							

13 The !Xun constituent order variation applies only with certain verbs. Also note that (17a) and (17b) have slightly different forms of the 1SG subject marker, though the nature of this alternation is not clear from the source.

(22) Skolt Saami; intra-group (Feist, 2015: 282)¹⁴

- a. *mii leäi ee'kk [ko pue'ðid tääzz*
 what be.PST.3SG year **when** come.PST.2PL PROX.SG.ILL
Če'vetjaurra?]
 place.name.ILL
 'What year was it, when you came here to Sevettijärvi?'

- b. [*ku pue'ttve'ted kuâssa*] *niõðstad lij â'lgg*
when come.PRS.2PL on.visit daughter.LOC.2PL be.PRS.3SG son
 'When you come to visit, your daughter will have a son.'

(23) Mongsen Ao; intra-group (Coupe, 2007: 202, 222)

- a. [*mətfatshəŋ nə tsəŋ-pà?*] *a-úk sə*
 person.name AGT spear-NMLZ NRL-pig ANAPH
 'the pig that Mechtseng speared'

- b. *nì a-ki [ípá? mətəm i tʃhà-tʃət-pà?]*
 1SG NRL-house EMPH manner PROX make-ABIL-NMLZ
 'my house that I am able to build like this here'

There are also some order variables (N = 36) that involve the positioning of affixes and clitics. Most of these involve variation in the position of an affix within a word, as seen in (24) and (25). But there are also some that involve an affix or clitic that attaches at variable positions in the phrase, for example in Tundra Nenets where in some contexts person agreement can be hosted by either verb or noun, as in (26).

(24) Bantawa; intra-group (Doornenbal, 2009: 274)

- a. *k^him k^har-a-k^ha-ci* b. *k^him k^har-a-ci-k^ha*
 house go-PST-see-DU
 'You two go home please!'

(25) Urarina; intra-group (Olawsky, 2006: 480, 524)

- a. *itçau-ru-rehete-lu* b. *itçau-rehete-kure-lu*
 live-PL-HAB-REM
 'They used to live.'

14 Some elements of the glossing are simplified in (22).

- (26) Tundra Nenets; intra-group (Nikolaeva, 2014: 323,329)¹⁵
- a. *yil'e-qm'a* *mərin'i*
live-PFV.AN city.PL.1SG
'the cities where I lived'

b. *to-qma-m'i* *yal'a-r'i-x^ona*
come-PFV.AN-1SG day-LIM-LOC
'on the same day when I came'

Only a minority of order variables are dialectal, but 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, as demonstrated in (27).¹⁶ In Turung, speakers in some villages sometimes use a different basic constituent order from those in others, as in (28). This is an instance where most dialects have a fixed order (SOV) while some villages show variation SOV ~ SVO, which is presumed to have arisen from contact with a neighboring SVO language (Morey, 2010: 513). Like the English double-object dative example in (16), the Turung and Ma'di examples show that dialects can sometimes be differentiated by word ordering. But this is rare compared to dialectal form variables.

- (27) Ma'di; Lokai vs. 'Burolo dialects (Blackings and Fabb, 2003: 176)
- a. *àmá* *èbī* *jā*
S O V
1PL.EXCL fish eat.NPST
'We are eating fish.'

b. *àmá jā èbī*
S V O

- (28) Turung; Tai villages vs. others (Morey, 2010: 511, 513)
- a. *phan n-gja* *phan n-gja* *la numsa*
S V O
type NEG-good type NEG-good take girl
'A bad (ghost) took a girl.'

15

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 and Luis, 2012: 220), in our coding we simply follow the authors of grammars in how they distinguish clitics vs. affixes.

16

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 constituents 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 (Blackings and Fabb, 2003: 541).

- b. *srownɡ wa [chkhɪ phe] [saa mkau]*
 S O V
 tiger DEF deer A.AG eat discard
 ‘The tiger ate the deer.’

Closer inspection of dialectal word-order variables reveals that they are noticeably concentrated in a small number of languages. Almost half the instances (15 of 33) are accounted for by two languages, Basque and !Xun, exemplified in (29) and (30), respectively. 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; western vs. other dialects (Hualde and Ortiz de Urbina, 2003: 524–525)

- a. *badakit etorriko ez dena*
 know come.FUT NEG AUX.COMP.DET
 ‘I know that he will not come.’

- b. *badakit ez direla etorriko*
 know NEG AUX.that come.FUT
 ‘I know that they will not come.’

- (30) !Xun; western W2 vs. other dialects (Heine and König, 2015: 92)¹⁷

- a. *ħa má ú kē m̃*
 he TOP go PST eat
 ‘He ate while going.’

- b. *ħa m é n!hō !hún g!!hōē*
 he TOP PST hit kill.SG dog
 ‘He beat the dog dead.’

Among the remaining nine instances of dialectal word-order variables, spread out over seven languages, a common factor is that contact with an unrelated

¹⁷ The order in (30a) is reported to be limited to the W2 dialect, where it occurs alongside order (30b), which is also found in all other dialects. Both examples here are from W2, since these most clearly show the order variation.

language is mentioned as a causal factor. This was already illustrated above for Turung, in (28), and another example is in the Jerusalem dialect of Domari, where possessor-possessum ordering is reported to have inverted due to intensive bilingualism with Arabic:

(31) Domari possession; Jerusalem vs. Syrian dialects (Matras, 2012: 168)

- | | |
|------------------------------------|------------------------------|
| a. <i>kury-os</i> <i>bɔy-im-ki</i> | b. <i>bɔy-im</i> <i>kuri</i> |
| house-3SG father-1SG.OBL-ABL | father-1SG.OBL house |
| ‘my father’s house’ | ‘my father’s house’ |

There are also a few instances of dialectal variation in affix order ($N = 6$). Examples are illustrated here from Slave negated verbs with incorporated postpositions, in (32), and the Biniñ Gun-wok immediacy affix, in (33). An example of dialectal order variation in clitics is found in Somali, where interrogatives may host pronominal and negative clitics in either order, as seen in (34).

(32) Slave; Hare vs. Bearlake (Rice, 1989: 777)¹⁸

- | | |
|--------------------------|--------------------------|
| a. <i>du-be-godihsho</i> | b. <i>be-du-godihsho</i> |
| NEG-3SG-forget | |
| ‘I forgot it.’ | |

(33) Biniñ Gun-wok; Gun-djeihmi vs. Kunwinjku clan lects (Evans, 2003: 320)

- | | |
|-----------------------------|-----------------------------|
| a. <i>ka-birri-h-karrme</i> | b. <i>ka-h-birri-karrme</i> |
| NPST-3AUG-IMM-have.NPST | |
| ‘Now they have it.’ | |

(34) Somali; central region vs. others (Saeed, 1999: 180, 275)

- | | |
|----------------------|----------------------|
| a. <i>ma=aad=áan</i> | b. <i>ma=áan=aad</i> |
| Q=2SG=NEG | |
| ‘not you?’ | |

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 or absence of a grammatical marker. A well-studied omission variable is in French verbal negation, where the particle *ne* is variably present or absent, as illustrated in (35). The single-marked version has over time become

¹⁸ Some aspects of the glossing are simplified in (32).

dominant in speech, and spread across geographic dialects, while the double-marked version remains in writing (Ashby, 1981; Armstrong, 2002; Martineau and Mougeon, 2003).

(35) French negation; written vs. spoken (Ashby, 1981)

je (ne) sais pas
 1SG NEG know NEG
 'I don't know'

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

The concept of coding (a)symmetry has often been applied to *contrastive* grammatical meanings, such as asymmetrical markedness relations in singular vs. plural, but it can also be applied to our structural types of variation in expressing the *same* meanings.¹⁹ While form variables are symmetrical differences between two alternant expressions, an omission variable is an asymmetrical difference. Just as coding asymmetries are thought to be governed by principles of efficient communication (Haspelmath, 2021), we might expect that omission variables should be governed primarily by informational efficiency, instead of social signaling.

Omission variables are less frequent than form variables, but more frequent than order variables. They account for 23% (N = 261) of all grammatical variables we identified, and at least one omission variable was identified for 40 of the 42 languages. Of the omission variables, 26% (N = 67) 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 article—in example (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.

¹⁹ We thank a reviewer for pointing out this connection to (a)symmetrical coding.

(36) Tundra Nenets; intra-group (Nikolaeva, 2014: 174)

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

Tundra Nenets also provides one of the few examples of an omission variable with a dialectal association. Negative clauses always begin with a negative particle, but speakers from the eastern region additionally use a *-q* "connegative" suffix on the verb, which is often (but not always) omitted by speakers from the western region.

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 better understand the behavior of structural types with respect to dialect differentiation by breaking our data down by degree of social contact between groups. Figures 4a and 4b illustrate grammatical variables, colored to distinguish dialectal variables in pink and intra-group variables in dark blue, grouped by degrees of social distance. Figure 4a shows raw count data, and 4b shows percentages of dialectal vs. intra-group. Figure 4b shows that around 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.

To test for a relationship between structural types, social distance, and dialectal differentiation, we fitted a mixed-effects regression model as follows.²⁰ The outcome to be predicted is whether a grammatical variable is dialectal or not, and the fixed effects are structural type and social distance. We modeled the fixed effects using treatment coding (or "dummy coding"), with form variables in situations of close contact as the baseline or "reference levels." The model then estimates the effect of switching to either an order or omission variable, the effect of increasing social distance, and finally an interaction effect of both switching type and increasing distance. Order and omission are coded as treatment contrasts, each being compared against form variables. Distance

20 R code and data used for this regression are available at <https://doi.org/10.5281/zenodo.7306236>.

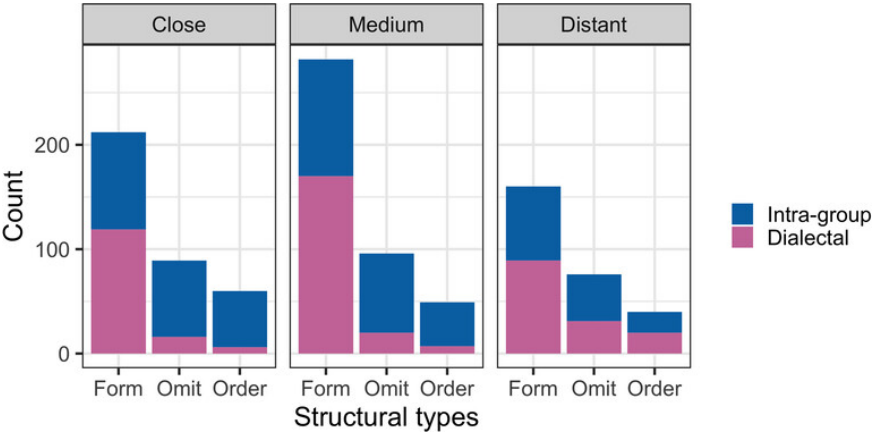


FIGURE 4A Counts of grammatical variables categorized as intra-group or dialectal, grouped by degrees of social contact

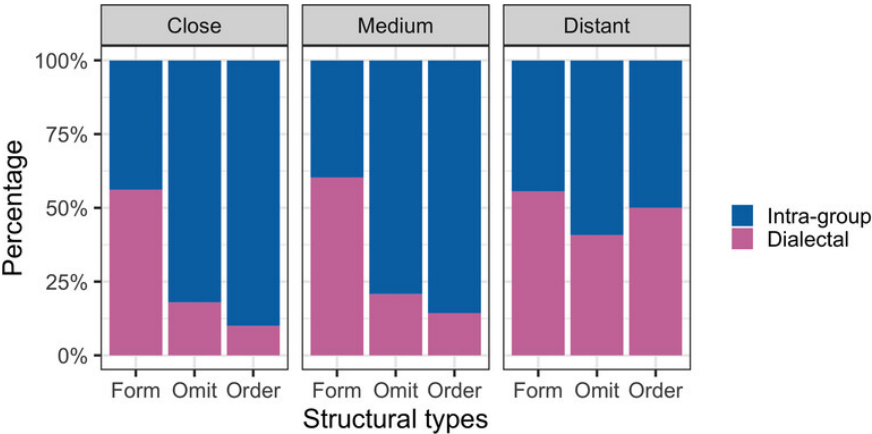


FIGURE 4B Percentages of intra-group vs. dialectal grammatical variables, grouped by degrees of social contact

is coded as a polynomial contrast, testing for a linear or quadratic 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 (using maximal language families as annotated in Glottolog [Hammarström et al., 2022]).²¹ As noted above, our data contains different quantities of data from different language families, but we control for this imbalance by including a random intercept for each family, and a random slope parameter for structural type in each family.²²

The model was fitted in R using the lmer package (Bates et al., 2015), with estimates of the predictors shown in Table 2. The intercept represents the log odds of a form variable, in a close-contact situation, being dialectal. This is not significantly different from zero; that is, there are even chances of it being dialectal or not. The fixed effects conform to our impressionistic analysis: comparing omission or order types to the form baseline (while keeping close contact as a reference level) produces highly significant, negative effects on the probability of a grammatical expression being dialectal. Meanwhile the effect of social distance, when considered with the form type as a reference level, is not significant.²³ But when considering the interaction of structural type with social distance, we find that greater social distance significantly increases the probability of an order variable being dialectal. For omission variables, the social distance effect is not significant, though it trends towards an increased probability of being dialectal.

It should be noted that under the treatment coding scheme used here, the negative effect of omission and order variable types is not a general effect, but rather one that holds under the condition of close social distance. The interaction between social distance and variable type suggests that at greater social

21 Modeling individual languages produces very similar results, as most of our families are represented by a single language. Modeling macroarea (Dryer, 1989) as a random effect also produces similar results, though with slightly larger effect-sizes and higher levels of statistical significance. We prefer the more conservative model using language families.

22 The random slope parameter causes some convergence issues when running the model, though it does converge after additional iterations, returning a “singular fit” warning that indicates zero variance for some random effects (Bates et al., 2015). These issues likely reflect the fact that some language families in the data have only a single data point for some structural types (e.g., Arawa has only a single order variable). We do not consider these warnings to have any impact on our results, as almost identical results are returned when we remove random slopes from the model (see supplementary materials, Section G, for alternative models). We nonetheless retain random slopes in the model reported in the main text, to demonstrate that our results are independent of individual language families.

23 lmer provides both linear and quadratic estimates of a three-level polynomial contrast. We here report whichever is the strongest estimate for each social distance factor.

TABLE 2 Model coefficients of a mixed-effects regression predicting dialectal status

Parameter	Reference levels for parameter estimate	Estimate and 95% CI (in log odds)	<i>p</i> value
Intercept	Form type in close contact	−0.06 [−0.55, 0.43]	<i>p</i> = .82
Type: Omission	Close contact	−1.96 [−2.54, −1.37]	<i>p</i> < .001
Type: Order	Close contact	−1.83 [−2.54, −1.12]	<i>p</i> < .001
Social distance (linear)	Form type	−0.23 [−0.72, 0.26]	<i>p</i> = .36
Social distance (quadratic)	–	0.73 [−0.02, 1.49]	<i>p</i> = .06
x Omission			
Social distance (linear)	–	1.39 [0.46, 2.32]	<i>p</i> < .01
x Order			

distances, order variables (and perhaps omission variables) become more like form variables in having a higher probability of distinguishing dialects. To further evaluate the significance of the interaction between social distance and structural type, we used ANOVA to compare the model in Table 2 with a similar model that does not include the interaction term. This shows that the interaction term reduces the deviance of the model (from 1,128 to 1,115) and improves the Akaike Information Criterion (from 1,150 to 1,145). The difference between models is statistically significant (*p* < .01).

The family-level intercepts for language families range from −2.02 (Baining) to 2.09 (Athapaskan), with a standard deviation of 1.16. 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 there to be differences between grammar writers in how much attention they pay to dialectology, and we suspect that these random intercepts most likely reflect grammar-writing methodologies, rather than actual differences between language families. The family-level slopes for the order type (compared to form) range from −2.20 (Austronesian) to −1.37 (Athapaskan), with a standard deviation of 0.43. The family-level slopes for the omission type range from −2.64 (Baining) to −1.11 (Athapaskan), with a standard deviation of 0.48. Notice that these family-level slopes are all negative, 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 crosslinguistic pattern, rather than being unduly influenced by exceptional families in our data.

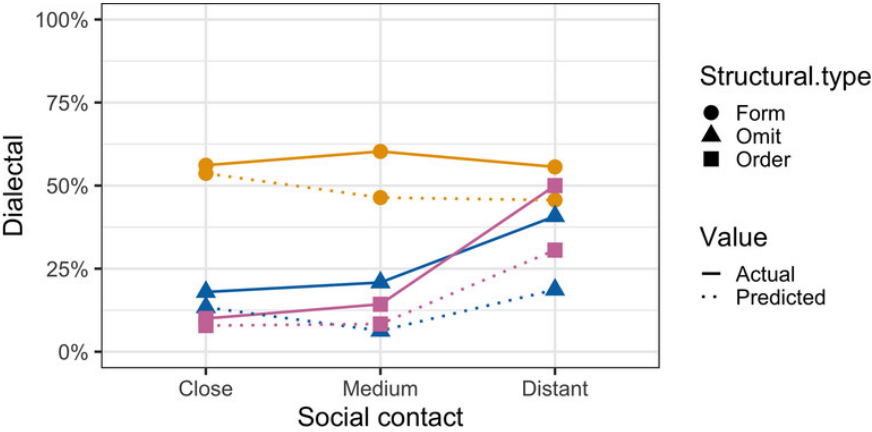


FIGURE 5 Actual data versus probabilities predicted by the model

To further evaluate the model we compared its predictions against the actual data. Figure 5 shows the actual percentages of dialectality in the data (as in Fig. 4b above), compared against the model’s predicted probabilities of a grammatical variable being dialectal. We use shapes to represent different structural types, and line types to represent actual data versus predictions. As the figure shows, the model is a fairly good fit for the data, though it does not predict the sharp increase in dialectal order variables that is found in the data. This is likely because a substantial number of the dialectal order variables, with distant social contact, come from just a few families (especially Basque and Kxa), and the model attributes this to family-level random effects rather than a general effect. Nonetheless, even when controlling for language family, the model still predicts that omission and order variables are more likely to be dialectal as social contact decreases.²⁴

6 Summary of findings and implications

One simple finding of our study is that grammatical markers often differentiate dialects. This supports the notion that social signaling can be integrated

24 The model also predicts a slight *decrease* in dialectality for form variables as dialects become more socially distant. This does not seem very plausible, since it would imply that dialects somehow become more similar in their grammatical markers as they lose social contact. We assume this is either an artifact of nonsignificant interaction effects in the model, or perhaps an artifact of grammar writers being less alert to the dialectal status of variables when the relevant dialects are not in close contact.

into the grammatical system, at least if we consider “surface forms” to be part of the grammar (cf. Cheshire, 1987; Labov, 1993). When reference grammars report two distinct markers as expressing the same grammatical meaning, in roughly half the instances this is reported to be a dialectal distinction. Although reference grammars cannot be read as comprehensive sources on dialectology, this finding nonetheless suggests that dialect differentiation is frequently intertwined with grammar. Whether this is simply a matter of “surface forms,” or whether it also results in grammatico-semantic differentiation, is an important question for further research (see Section 4.1).

Furthermore, our survey reveals that different structural types of variation exhibit different patterns with respect to dialectality. Dialect differentiation by form variables applies equally under 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 form differentiation is established during periods of close contact, and little is added once they move apart. By contrast, order and omission variables are rarely dialectal in situations of close contact, but order variables become more likely to differentiate dialects as they become distant from one another. Since social signaling is only relevant to the extent that groups are in social contact, these findings suggest that form variables are driven to a greater extent by social signaling, compared to order and omission variables. This is compatible with studies of language contact that identify divergence of “matter” or “lexicon” alongside convergence of “pattern” or “structure.” Our study builds on the previous research by operationalizing specific types of linguistic variation and showing systematic differences between them in a crosslinguistic sample.

6.1 *Language speciation and diversification*

In the introduction we defined “linguistic divergence” as diversification driven by social contact as opposed to social separation. 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 6 extends the schema from Fig. 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 (dotted lines) between these groups 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, performing much of the initial work in “language speciation.” Form differences 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.

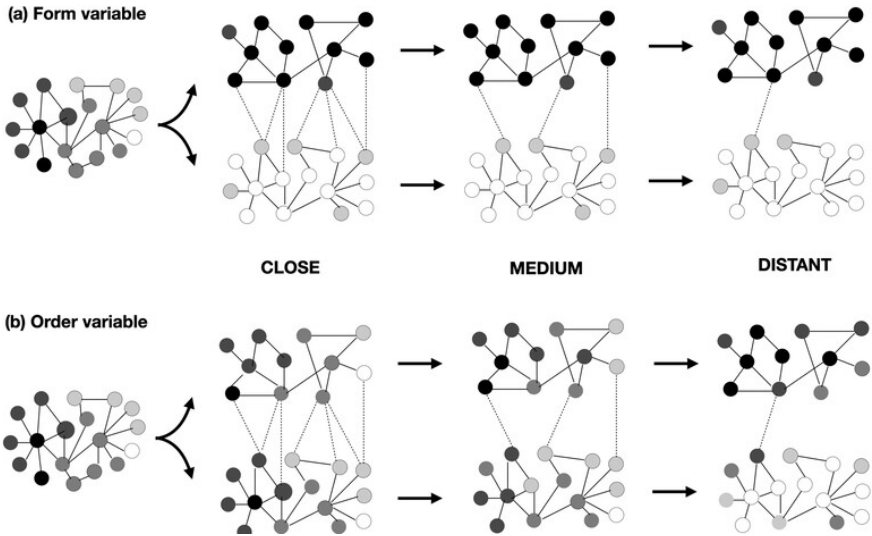


FIGURE 6 Schematic model of how form and order variables differentiate dialects over time

If we project this schema onto a multi-millennial timescale, it would suggest that language families diversifying “in situ,” with prolonged social contact between related varieties, should exhibit more diversity in the forms of grammatical markers compared to families that diversify in a more dispersed manner. Differentiation of grammatical markers (and lexicon) could become so extensive that these varieties would become quite distinct languages, despite a lack of geographic or social separation (François, 2011; François, 2012). 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 favors more diversification of the lexicon. This is compatible with the findings of the current study, though our findings suggest that social contact may favor diversification not just in open-class lexical items but also in grammatical markers.

If social signaling is more easily achieved with grammatical markers than with linear ordering, this is also consistent with sociolinguistic research. There are many well-studied examples of form variables that are salient markers of social and regional identity, such as *isn't* ~ *ain't* in English and *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*.

6.2 *Why do order variables resist social signaling?*

Why exactly should order variation exhibit less social signaling, 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 and Gordon, 2003: 187; Cheshire et al., 2005). One possible explanation for their lack of social signaling is that these strong language-internal factors inhibit the development of social signaling. 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 social signaling.

As an example of a linguistic variable that is strongly predicted by linguistic contextual factors, take the standard English dative alternation (Bresnan and Nikitina, 2009; Bresnan and Ford, 2010). Regression modeling 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. A model combining these predictors achieves 94.5% accuracy on unseen corpus data (Bresnan and 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 a social signaling function, and this may be precisely because there is so little variance left over after linguistic context is factored out (but see Jensen et al., 2018). 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 and Jaeger, 2015).

The argument outlined above is similar to a theory of social signaling in terms of expectations and surprisal (Rácz, 2013; Jaeger and 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 a social signal, it should have a high surprisal (negative log probability) based on purely linguistic context. For example, in British English, glottalization of stops has low surprisal in coda position, where it occurs quite frequently as a function of language-internal articulatory patterns, and therefore has little potential to be interpreted as a social signal. But in intervocalic position it has higher surprisal, facilitating social signaling by intervocalic t-glottalization (Rácz, 2013: 145). A theory of social signaling in terms of surprisal is compatible with the idea that order variables resist social signaling because they are so heavily conditioned by linguistic context. Again,

this would imply that variant selection leaves little residual surprisal, and surprisal is key to the interpretation of social signaling.

Acknowledgments

We are grateful to the speakers of 42 languages who shared the knowledge that forms the basis for this study, and the field linguists who collated this into reference grammars. We are also grateful for the insightful comments of many colleagues: Brett Baker, Lucy Davidson, Rebecca Defina, Chloe Diskin-Holdaway, Gabriela Garrido, Stephen Levinson, Jonathon Lum, Luisa Miceli, Rachel Nordlinger, Catherine Travis, Jill Vaughan, James Walker, and two anonymous reviewers. This research was funded by the Australian Research Council, grant DE180100872.

Supplementary materials

In the supplementary materials, we include further information on the selection of reference grammars and their bibliographic details. We also describe the process for identifying grammatical variables and how they were categorised into structural types, and provide more information about the coding of social distance, and our tests for inter-coder reliability. Finally, we present an alternative regression model without random slopes for language family by structural type, which provides very similar results to the model presented in the main paper. The supplementary materials can be accessed here: <https://doi.org/10.6084/m9.figshare.22666360>.

References

The complete list of reference grammar sources can be found in the supplementary materials.

- Abraham, Werner. 2006. Dialect and typology: Where they meet—and where they don't. In Terttu Nevalainen, Juhani Klemola, and Mikko Laitinen (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* 23(3): 231–273. [https://doi.org/10.1016/S0271-5309\(03\)00012-0](https://doi.org/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. [https://doi.org/10.1016/S0388-0001\(01\)00015-8](https://doi.org/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. <https://doi.org/10.2307/414345>.
- Atkinson, Quentin D., Andrew Meade, Chris Venditti, Simon J. Greenhill, and Mark Pagel. 2008. Languages evolve in punctuational bursts. *Science* 319(5863). <https://doi.org/10.1126/science.1149683>.
- Bates, Douglas, Martin Maechler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1): 1–48.
- Bateson, Gregory. 1935. Culture contact and schismogenesis. *Man* 35: 178–183. <https://doi.org/10.2307/2789408>.
- Blackings, Mairi and Nigel Fabb. 2003. *A Grammar of Ma'di*. Berlin: Mouton de Gruyter.
- Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.
- Blythe, Richard A. and William Croft. 2021. How individuals change language. *PLOS ONE* 16(6). e0252582. <https://doi.org/10.1371/journal.pone.0252582>.
- Boye, Kasper and Peter Harder. 2012. A usage-based theory of grammatical status and grammaticalization. *Language* 88(1): 1–44.
- Bresnan, Joan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1): 168–213.
- Bresnan, Joan and Tatiana Nikitina. 2009. The gradience of the dative alternation. In Linda Uyechi and Lian Hee Wee (eds.), *Reality Exploration and Discovery: Pattern Interaction in Language and Life*, 161–184. Stanford: CSLI.
- Broadwell, George Aaron. 2006. *A Choctaw Reference Grammar*. Lincoln: University of Nebraska Press.
- Campbell-Kibler, Kathryn. 2010. The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change* 22(3): 423–441. <https://doi.org/10.1017/S0954394510000177>.
- Campbell-Kibler, Kathryn. 2016. Towards a cognitively realistic model of meaningful sociolinguistic variation. In Anna M. Babel (ed.), *Awareness and Control in Sociolinguistic Research*, 123–151. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139680448.008>.
- Chambers, Jack and Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge: Cambridge University Press.
- Cheshire, Jenny. 1987. Syntactic variation, the linguistic variable, and sociolinguistic theory. *Linguistics* 25(2): 257–282. <https://doi.org/10.1515/ling.1987.25.2.257>.
- Cheshire, Jenny, Paul Kerswill, and Ann Williams. 2005. Phonology, grammar, and discourse in dialect convergence. In Peter Auer, Frans Hinskens, and Paul Kerswill (eds.), *Dialect Change: Convergence and Divergence in European Languages*, 135–168. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511486623.007>.

- Coupe, Alexander R. 2007. *A Grammar of Mongsen Ao*. Berlin: De Gruyter.
- Croft, William. 2000. *Explaining Language Change: An Evolutionary Approach*. London: Longman.
- 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.
- De Sutter, Gert. 2005. *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen*. PhD thesis, KU Leuven.
- Dediu, Dan, Michael Cysouw, Stephen C. Levinson, Andrea Baronchelli, Morten H. Christiansen, William Croft, Nicholas Evans, Simon Garrod, Russell D. Gray, Anne Kandler, and Elena Lieven. 2013. Cultural evolution of language. In Peter J. Richerson and Morten H. Christiansen (eds.), *Cultural Evolution: Society, Technology, Language, and Religion*, 303–332. Cambridge: 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* 2018(254): 139–163. <https://doi.org/10.1515/ijsl-2018-0037>.
- Di Garbo, Francesca, Eri Kashima, Ricardo Napoleão de Souza, and Kaius Sinnemäki. 2021. Concepts and methods for integrating language typology and sociolinguistics. In Silvia Ballarè and Guglielmo Inglese (eds.), *Tipologia e sociolinguistica: Verso un approccio integrato allo studio della variazione*, 143–176. Milano: Officinaventuno. <https://doi.org/10.17469/O2105SLI000005>.
- Díaz Collazos, Ana María. 2015. *Desarrollo sociolingüístico del voseo en la región andina de Colombia (1555–1976)*. Berlin: De Gruyter.
- 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. <https://doi.org/10.5281/zenodo.1477799>.
- Doornenbal, Marius. 2009. *A Grammar of Bantawa*. Meteren: Netherlands Graduate School of Linguistics.
- Drager, Katie and M. Joelle Kirtley. 2016. Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. In Anna M. Babel (ed.), *Awareness and Control in Sociolinguistic Research*, 1–24. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139680448.003>.
- Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2): 257–292. <https://doi.org/10.1075/sl.13.2.03dry>.
- Dryer, Matthew S. and Martin Haspelmath (eds.). 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://wals.info/> (accessed November 4, 2022).
- Dunbar, Robin I.M. 2003. The origin and subsequent evolution of language. *Language*

- Evolution*, 219–234. Oxford: Oxford University Press. <https://doi.org/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. <https://doi.org/10.1353/lan.2019.0072>.
- Ellison, T. Mark and Luisa Miceli. 2017. Language monitoring in bilinguals as a mechanism for rapid lexical divergence. *Language* 93(2): 255–287. <https://doi.org/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 Milly Crevels and Pieter Muysken (eds.), *Language Dispersal, Diversification, and Contact: A Global Perspective*. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780198723813.003.0016>.
- Errington, J. Joseph. 1985. On the nature of the sociolinguistic sign: Describing the Javanese speech levels. In Elizabeth Mertz and Richard J. Parmentier (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 Michela Cennamo and Claudio Fabrizio (eds.), *Historical Linguistics, 2015: Selected Papers from the 22nd International Conference on Historical Linguistics*, 564–591. Amsterdam: John Benjamins.
- 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*. PhD thesis, Australian National University.
- François, Alexandre. 2011. Social ecology and language history in the northern Vanuatu linkage: A tale of divergence and convergence. *Journal of Historical Linguistics* 1(2): 175–246. <https://doi.org/10.1075/jhl.1.2.03fra>.
- François, Alexandre. 2012. The dynamics of linguistic diversity: Egalitarian multilingualism and power imbalance among northern Vanuatu languages. *International Journal of the Sociology of Language* 214: 85–110.
- François, Alexandre. 2014. Trees, waves and linkages: Models of language diversification. In Claire Bower and Bethwyn Evans (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 Ilana Mushin and Brett Baker (eds.), *Discourse and Grammar in Australian Languages*, 111–134. Amsterdam: John Benjamins.

- Gal, Susan. 2016. Sociolinguistic differentiation. In Nikolas Coupland (ed.), *Sociolinguistics: Theoretical Debates*, 113–136. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781107449787.006>.
- Gardani, Francesco. 2020. Borrowing matter and pattern in morphology: An overview. *Morphology* 30: 263–282. <https://doi.org/10.1007/s11525-020-09371-5>.
- 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 Peter Forster and Colin Renfrew (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* 17(3): 445–463. <https://doi.org/10.1017/S1360674313000117>.
- Grace, George W. 1981. *An Essay on Language*. Columbia, SC: Hornbeam Press.
- Greenhill, Simon J., Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C. Levinson, and Russell D. Gray. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences* 114(42). E8822–E8829. <https://doi.org/10.1073/pnas.1700388114>.
- Gumperz, John J. and Robert Wilson. 1971. Convergence and creolization: A case from the Indo-Aryan/Dravidian border in India. In Dell Hymes (ed.), *Pidginization and Creolization of Languages*, 151–167. Cambridge: Cambridge University Press.
- Haddican, Bill, Daniel Ezra Johnson, Joel Wallenberg, and Anders Holmberg. 2020. Variation and change in the particle verb alternation across English dialects. In Karen V. Beaman, Isabelle Buchstaller, Sue Fox, and James A. Walker (eds.), *Advancing Socio-grammatical Variation and Change*, 15–31. London: Routledge. <https://doi.org/10.4324/9780429282720-14>.
- Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2022. *Glottolog* 4.6. Leipzig. <https://doi.org/10.5281/zenodo.6578297>.
- Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1): 31–80.
- Haspelmath, Martin. 2021. Explaining grammatical coding asymmetries: Form—frequency correspondences and predictability. *Journal of Linguistics* 57(3): 605–633. <https://doi.org/10.1017/S0022226720000535>.
- Haugen, Einar. 1988. Dialects as stepping stones to a language. In Alan R. Thomas (ed.), *Methods in Dialectology*, 666–673. Clevedon: Multilingual Matters.
- Heine, Bernd and Christa König. 2015. *The !Xun Language: A Dialect Grammar of Northern Khoisan*. Cologne: Rüdiger Köppe.
- Hellwig, Birgit. 2019. *A Grammar of Qa̱et*. Berlin: Mouton de Gruyter.

- Himmelmann, Nikolaus P. 2022. Prosodic phrasing and the emergence of phrase structure. *Linguistics* 60(3): 715–743. <https://doi.org/10.1515/ling-2020-0135>.
- Hinskens, Frans. 1998. Variation studies in dialectology and three types of sound change. *Sociolinguistica* 12(1998): 155–193. <https://doi.org/10.1515/9783110245172.155>.
- Hopper, Paul J. and Elizabeth Closs Traugott. 2003. *Grammaticalization*. 2nd edn. Cambridge: Cambridge University Press.
- Hualde, José Ignacio and Jon Ortiz de Urbina. 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. <https://doi.org/10.1016/j.cogpsych.2010.02.002>.
- Jaeger, T. Florian and Kodi Weatherholtz. 2016. What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology* 7. <https://doi.org/10.3389/fpsyg.2016.01115>.
- Jahr, Ernst Håkon. 2017. A pioneering sociolinguistic concept: Amund B. Larsen and his discovery in the 1880s of neighbour opposition as a socio-psychological mechanism in linguistic change—and its rediscovery in the 1980s as hyperdialectism. *European Journal of Scandinavian Studies* 47(2): 308–319. <https://doi.org/10.1515/ejss-2017-0020>.
- Jenset, Gard B., Barbara McGillivray, and Michael Rundell. 2018. The dative alternation revisited: Fresh insights from contemporary British spoken data. *Corpus Approaches to Contemporary British Speech*, 185–207. London: Routledge.
- Kerswill, Paul and Ann Williams. 2002. “Salience” as an explanatory factor in language change: Evidence from dialect levelling in urban England. In Mari C. Jones and Edith Esch (eds.), *Language Change: The Interplay of Internal, External and Extra-linguistic Factors*, 81–110. Berlin: Mouton de Gruyter.
- Kurumada, Chigusa and T. Florian Jaeger. 2015. Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83: 152–178. <https://doi.org/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. NWAV 22, University of Ottawa, October.
- Lai, Wei, Péter Rácz, and Gareth Roberts. 2020. Experience with a linguistic variant affects the acquisition of its sociolinguistic meaning: An alien-language-learning experiment. *Cognitive Science* 44(4). e12832. <https://doi.org/10.1111/cogs.12832>.
- Lavandera, Beatriz R. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7(2): 171–182. <https://doi.org/10.1017/S0047404500005510>.
- Levinson, Stephen C. 1988. Conceptual problems in the study of regional and cultural style. In Norbert Dittmar and Peter Schlobinski (eds.), *The Sociolinguistics of Urban Vernaculars*, 161–190. Berlin: Mouton de Gruyter.

- Lehmann, Christian. 1995. *Thoughts on Grammaticalization*. Munich: Lincom Europa.
- Levon, Erez and Isabelle Buchstaller. 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. <https://doi.org/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 and James N. Stanford. 2017. Documenting sociolinguistic variation in lesser-studied indigenous communities: Practical methods and solutions. In Kristine A. Hildebrandt, Carmen Jany, and Wilson Silva (eds.), *Documenting Variation in Endangered Languages*, 116–136. Honolulu: University of Hawai'i Press.
- Martineau, France and Raymond Mougeon. 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.
- Matras, Yaron and Jeanette Sakel. 2007. Investigating the mechanisms of pattern replication in language convergence: *Studies in Language* 31(4): 829–865. <https://doi.org/10.1075/sl.31.4.05mat>.
- Matsumae, Hiromi, Peter Ranacher, Patrick E. Savage, Damián E. Blasi, Thomas E. Currie, Kae Koganebuchi, Nao Nishida, Takehiro Sato, Hideyuki Tanabe, Atsushi Tajima, Steven Brown, Mark Stoneking, Kentaro K. Shimizu, Hiroki Oota, and Balthasar Bickel. 2021. Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Science Advances* 7(34): eabd9223. <https://doi.org/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 Francesco Gardani, Peter Arkadiev, and Nino Amiridze (eds.), *Borrowed Morphology*, 189–218. Berlin: Mouton de Gruyter.
- Meyerhoff, Miriam and James A. Walker. 2012. Grammatical variation in Bequia (St Vincent and the Grenadines). *Journal of Pidgin and Creole Languages* 27: 209–234.
- Miceli, Luisa, T. Mark Ellison, Bethwyn Evans, and Simon J. Greenhill. 2016. Can we identify bilingual-led lexical differentiation in Oceanic? Australian Linguistic Society Conference, Monash University, December 7.
- Milroy, Lesley and Matthew Gordon. 2003. *Sociolinguistics: Method and Interpretation*. 2nd edn. 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:ngu language of north-east Arnhem Land. *Canberra Anthropology* 1: 51–60.
- Nettle, Daniel and Robin I.M. Dunbar. 1997. Social markers and the evolution of reciprocal exchange. *Current Anthropology* 38(1): 93–99. <https://doi.org/10.1086/204588>.

- Nikolaeva, Irina. 2014. *A Grammar of Tundra Nenets*. Berlin: De Gruyter Mouton.
- Olawsky, Knut J. 2006. *A Grammar of Urarina*. Berlin: Mouton de Gruyter.
- Paul, Hermann. 1888. *Principles of the History of Language*. London: Swan Sonnenschein, Lowrey & Co.
- 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*. Berlin: De Gruyter Mouton.
- 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 Bruno Galantucci and Simon Garrod (eds.), *Experimental Semiotics: A New Approach for Studying the Emergence and the Evolution of Human Communication*, 138–159. Amsterdam: John Benjamins.
- 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 Mark Durie and Malcolm Ross (eds.), *The Comparative Method Reviewed*, 180–217. Oxford: Oxford University Press.
- Ross, Malcolm. 2001. Contact-induced change in Oceanic languages in north-west Melanesia. In Alexandra Y. Aikhenvald and R.M.W. Dixon (eds.), *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*, 134–166. Oxford: Oxford University Press.
- Röthlisberger, Melanie and Sali A. Tagliamonte. 2020. The social embedding of a syntactic alternation: Variable particle placement in Ontario English. *Language Variation and Change* 32(3): 317–348. <https://doi.org/10.1017/S0954394520000174>.
- Saeed, John. 1999. *Somali*. Amsterdam: John Benjamins.
- Saussure, Ferdinand de. 1959. *Course in General Linguistics*. New York: The Philosophical Society.
- Schad, Daniel J., Shravan Vasishth, Sven Hohenstein, and Reinhold Kliegl. 2020. How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language* 110: 104038. <https://doi.org/10.1016/j.jml.2019.104038>.
- Schwenter, Scott A. and Rena Torres Cacoullos. 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. <https://doi.org/10.1075/resla.27.2.13sch>.
- Shih, Stephanie S. and Kie Zuraw. 2017. Phonological conditions on variable adjective and noun word order in Tagalog. *Language: Phonological Data and Analysis* 93(4): e317–e352. <https://doi.org/10.1353/lan.2017.0075>.
- Siewierska, Anna and Willem B. Hollmann. 2007. Ditransitive clauses in English with special reference to Lancashire dialect. In Mike Hannay and Gerard J. Steen (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: 1–30.
- Silverstein, Michael. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & Communication* 23: 193–229.
- Smith, Ian and Johnson, Steve. 2000. Kugu Nganhcara. In R.M.W. Dixon and Barry J. Blake (eds.), *The Handbook of Australian Languages*, vol. 5, 355–490. Melbourne: Oxford University Press.
- Sneller, Betsy and Gareth Roberts. 2018. Why some behaviors spread while others don't: A laboratory simulation of dialect contact. *Cognition* 170: 298–311. <https://doi.org/10.1016/j.cognition.2017.10.014>.
- Spencer, Andrew and Ana R. Luis. 2012. *Clitics: An Introduction*. Cambridge: Cambridge University Press.
- Squires, Lauren. 2016. Processing grammatical differences: Perceiving versus noticing. In Anna M. Babel (ed.), *Awareness and Control in Sociolinguistic Research*, 80–103. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139680448.006>.
- 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.
- Tabouret-Keller, Andrée. 2017. Language and identity. In Florian Coulmas (ed.), *The Handbook of Sociolinguistics*, 315–326. John Wiley. <https://doi.org/10.1002/9781405166256.ch19>.
- Tamminga, Meredith, Laurel MacKenzie, and David Embick. 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, T. Florian Jaeger, and David Orr. 2011. Lexical variation in relativizer frequency. In H. Simon and H. Wiese (eds.), *Expecting the Unexpected: Exceptions in Grammar*, 175–195. Berlin: De Gruyter.
- Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. Empirical foundations for a theory of language change. In W. Lehmann and Y. Malkiel (eds.), *Directions for Historical Linguistics*, 95–198. Austin: University of Texas Press.
- Wichmann, Søren. 2019. How to distinguish languages and dialects. *Computational Linguistics* 45(4): 823–831. https://doi.org/10.1162/coli_a_00366.
- Wolfram, Walt. 1969. *A Sociolinguistic Description of Detroit Negro Speech*. Washington, DC: Center for Applied Linguistics.

- Zariquiey, Roberto. 2011. Aproximación dialectológica a la lengua cashibo-cacataibo (pano). *Lexis* 35(1): 5–46.
- Zariquiey, Roberto. 2018. *A Grammar of Kakataibo*. Berlin: Mouton de Gruyter.