Dialect differences and linguistic divergence: A cross-linguistic survey of grammatical variation

Supplementary Information

A. Selection of source materials

Grammars were selected with the aim of achieving geographic and genealogical diversity. However we faced some constraints on which languages could be included in the study, as we required (a) an available reference grammar, either present in our research library or freely available for download; (b) the grammar needed to be detailed enough to document variation. We started by searching through grammars in three major series, Mouton Grammar Library, Language Science Press Studies in Diversity Linguistics, and Pacific Linguistics. We selected grammars from language families or regions not yet represented in our sample, and included a grammar if it was found to have some information on variation. We also included grammars not in those series, when detailed grammars came to our attention that represented a family for which we had no sample. The languages and sources sampled in the study are listed in Table A1, with family groups as identified in Glottolog (Hammarström et al. 2022).

Table A1. Languages / families sampled and the sources used (see References list below)

FAMILY	LANGUAGE	GLOTTOCODE	SOURCE
Afro-Asiatic	Somali	soma1255	(Saeed 1999)
Algonquian	Nishnaabemwin	otta1242	(Valentine 2001)
Araucanian	Mapuche	mapu1245	(Smeets 2008)
Arawa	Kulina / Culina	culi1244	(Dienst 2014)
Athapaskan	Slave	slav1253	(Rice 1989)
Austroasiatic	Kharia	khar1287	(Peterson 2010)
Austroasiatic	Pnar	pnar1238	(Ring 2015)
Austronesian	Hoava	hoav1238	(Davis 2003)
Austronesian	Tigak	tiga1245	(Beaumont 1979)
Austronesian	Fijian	fiji1243	(Dixon 1988)
Austronesian	Madurese	nucl1460	(Davies 2010)
Baining	Qaqet	qaqe1238	(Hellwig 2019)
Basque	Basque	basq1248	(Hualde & Urbina 2003)
Gunwinyguan	Bininj Gun-wok	gunw1252	(Evans 2003)
Indo-European	Domari	doma1258	(Matras 2012)

Indo-European	Norman French	angl1258	(Liddicoat 1994)
Kxa	!Xun	juku1256	(Heine & König 2015)
Maningrida	Ndjébbana	djee1236	(McKay 2000)
Muskogean	Choctaw	choc1276	(Broadwell 2006)
Muskogean	Creek	cree1270	(Martin 2011)
Naduhup	Hup	hupd1244	(Epps 2008)
Nakh-Daghestanian	Sanzhi Dargwa	sanz1248	(Forker 2020)
Niger-Congo	Kisi	kiss1245	(Childs 2011)
Niger-Congo	Supyire	supy1237	(Carlson 2011)
Nilo-Saharan	Ma'di	madi1260	(Blackings & Fabb 2003)
Nyulnyulan	Yawuru	yawu1244	(Hosokawa 2011)
Pama-Nyungan	Kugu Nganhcara	pama1253	(Smith & Johnson 2000)
Pama-Nyungan	Ngiyambaa	wang1291	(Donaldson 1980)
Pama-Nyungan	Nyangumarta	nyan1301	(Sharp 2004)
Pano-Tacanan	Kakataibo	cash1251	(Zariquiey 2018)
Quechuan	Yauyos Quechua	yauy1235	(Shimelman 2016)
Sino-Tibetan	Mongsen Ao	mong1332	(Coupe 2007)
Sino-Tibetan	Turung	turu1249	(Morey 2010)
Sino-Tibetan	Bantawa	bant1281	(Doornenbal 2009)
Sino-Tibetan	Southern Min	minn1241	(Chen 2020)
Sino-Tibetan	Qiang	upst1234	(LaPolla & Huang 2003)
Tai-Kadai	Lao	laoo1244	(Enfield 2007)
Uralic	Skolt Saami	skol1241	(Feist 2015)
Uralic	Tundra Nenets	nene1249	(Nikolaeva 2014)
Urarina	Urarina	urar1246	(Olawsky 2006)
Western Daly	Emmi/Menhdhe	amem1235	(Ford 2011)
Yam	Komnzo	wara1294	(Döhler 2018)

This dataset provides a reasonably diverse sample of human languages, though it is ultimately a convenience sample, rather than a formally balanced or structured 'diversity sample' (Miestamo et al. 2016). The genus-macroarea metric proposed by Miestamo provides one way of evaluating how closely the number of genera in our sample reflects the number of genera in each continent-sized 'macroarea' (Dryer 1989; Hammarström & Donohue 2014). Table A2 shows that our sample includes multiple genera for all macroareas, and overall includes around 6% of all genera identified by Dryer (2013). Our sample is somewhat over-representative of the Australia and Eurasia macroareas, where we have sampled several genera from relatively small numbers of

total genera. By contrast, our sample is under-representative of both North America, and of Papunesia (a macroarea spanning Melanesia and the Pacific). Note that our sample has a small additional imbalance in that it includes multiple languages from two genera, namely Muskogean (two languages) and Oceanic (three languages). To control for these imbalances, our statistical method includes Glottolog maximal language families as random effects.

Table A2. Sample coverage of genera by macroarea

Macroarea	Genera	Genera in sample	Coverage
Africa	123	5	4.1%
Australia	40	7	17.5%
Eurasia	92	13	14.1%
North America	108	3	2.8%
Papunesia	167	4	2.4%
South America	112	6	5.4%
ALL	**625**	**38**	**6.1%**

B. Identifying grammatical variables

Coding of grammatical variables was done using a mixture of keyword searches, and reading either in PDF or in hard copy. The keyword method (see details below) is more systematic and replicable, and was applied to all reference grammars used in the study, with the exception of the Nishnaabemwin grammar (Valentine 2001).[1] The reading method is less systematic, but we used it as an 'auxiliary' method to identify further variables that were missed in keyword searches. It is difficult or impossible to read entire reference grammars (some up to 1000 pages long), with equal attention to all sections. Reading was used to various extents with various reference grammars, but a certain amount of reading was used for all grammars, as the keyword method only highlights relevant terms, but is not in itself sufficient to flesh out the basic characteristics of a grammatical variable.

[1] For Valentine (2001), we had access to a hard copy but no digital version. The grammar contains a dedicated section for documenting dialectal variables, and we took advantage of this as a short-cut to identifying what we would otherwise find using keyword searches.

Our keyword search method was performed in Adobe Acrobat (version 2021-011), using each of the following search-terms for each reference grammar:

*vari** (wildcard to match *variant, variation, variable,* etc.)

*option**

dialect

order

older, younger

*conservative, innovat**

casual, formal

slow, fast

Sections of text in which one or more keywords appear were then read, and if the description matched our definition of a grammatical variable, we added one or more records to our database (see coding conventions below). The keywords method provided an effective tool to identify a large number of linguistic variables in a wide range of languages, using a relatively systematic procedure. However we also included any other grammatical variables that we discovered by more general reading, as this increased the overall number of datapoints in the study, bringing it closer to a complete representation of the variables documented in the reference grammars, while also recognising that there is no practical method for representing *every* variable in the grammar.

In addition to the search-terms listed above, we also identified language-specific search-terms in each reference grammar. This was done by consulting the introductory chapter, which usually contains an overview of dialectology in the region under study. A few grammars (e.g. Urarina: Olawsky 2006) also had an additional chapter dedicated to sociolinguistic and/or dialectal variation. These sections and chapters typically yield important search-terms for identifying dialect differences in the rest of the text. For example in the Kakataibo grammar, the first chapter indicates that dialectal differences in the text will mostly refer to the San Alejandro, Aguatía and Sungaroyacu rivers (Zariquiey 2018: 34). These dialect labels are then used as additional search-terms for this text. In many grammars, the introductory statement on dialectology uses directional terms such as 'Eastern and Western' dialects, in which case these again are then used as additional search-terms. Reading of introductory sections on dialectology

thus formed another basic step in our coding method, and additionally informed our coding of social contact between dialect groups (see below).

C. Coding structural types

A grammatical variable was entered in the database wherever the text reports more than one way of expressing the same grammatical meaning or function. Each variable was coded as one of three main structural types, as described in section §4. To repeat those core definitions:

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

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

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

However there was also a residue of around 8% of variables which exhibit some mixture of the properties used to identify the main types. We labelled these 'Construction' variables. In Construction variables, each variant has some element that is not present in the other (otherwise it would be an Order or Omission variable) and there is also some difference in the linear configuration (otherwise it would be a Form variable).

For example, to express duality of a nominal in Yawuru, speakers may use either a dual suffix *-milidyarri* or a free numeral *kudyarra* 'two'.[2] These involve distinct morphemes but also (optionally) an order distinction, and the variable is therefore classified as Construction.

Yawuru: Dual marking

a. N-milidyarri

b. kudyarra N (~ N kudyarra) (Hosokawa 2011: 30)

Another example comes from Ndjébbana, where counterfactual statements may either use a counterfactual suffix directly attached to the verb (21a), or a non-finite main verb combined with an auxiliary that carries the counterfactual suffix (21b). The single-verb construction uses different counterfactual suffix forms from the auxiliary construction, so there is a form

[2] Another variant uses both markers (e.g. N-*milidyarri kudyarra*). This is an omission variable with respect to N-*milidyarri*.

alternation between grammatical markers, but also an ordering difference in how these markers are arranged with respect to the lexical verb. There is also an omission relationship, with the auxiliary verb present in one variant and not the other.

Ndjébbana

a. kóma nga-ya-rarraddja-ngóna (coastal clans)
NEG 1MIN.A-3MIN.O-clean-CF

b. kóma na-rórrddja nga-yangka-yína (riverine clans)
NEG INF-clean 1MIN.A>3MIN.O-do-CF
'I did not clean it.' (McKay 2000: 224)

A different type of example comes from the expression of the habitual past in Sanzhi Dargwa. This can be formed by adding the habitual past suffix *-a* and a 1st (*-di*) or 2nd (*-t:e*, *-t:a*) person agreement suffix to the imperfective form of the verb. In the third person, on the other hand, habitual past is marked with a single suffix *-i*. However this paradigm is optionally levelled, using *-i* alone to mark habitual past in any person with no change in meaning. Thus for 1st (or 2nd) person participants, there is variation between two possible verb forms, one of which uses two suffixes while the other uses a single suffix. The variable is therefore categorised as Construction (rather than e.g. Form).

Sanzhi Dargwa: Habitual paast

a. Nuš:a-l k:ač a-b-irq'-a-di
1PL-ERG touch NEG-NEUT-do.IPFV-HAB.PST-1

b. Nuš:a-l k:ač a-b-irq'-i
1PL-ERG touch NEG-NEUT-do.IPFV-HAB.PST.3
'We did not touch it' (Forker 2020: 246)

The Construction type also encompasses variation between segmentable and fused grammatical morphemes. For example, older Boumaa Fijian speakers fuse the numeral *dua* 'one' with the common noun article *a*, forming *daa*. Standard Fijian speakers, and younger speakers of the Boumaa dialect, produce the numeral and the article separately. This, and other cases of fusion, cannot be categorised into one of the main types because there is no meaningful change in order, and there are no morphemes being entirely omitted.

Fijian: Indefinite articles ‘one N, a N’

a.	*dua*	*a*	*N*	(Standard Fijian, and younger Boumaa speakers)
	one	Indef	N	
b.	*daa*		*N*	(Older Boumaa Fijian)
	one.Indef		N	(Dixon 1988: 146)

D. Other annotations

For each grammatical variable, we also annotated an abbreviated representation of its meaning or function (e.g. ‘FUT’, ‘NEG’, ‘N.PL’, ‘Pron.PL’, ‘clause.transitive’). We also used two columns to represent the two variable forms. Where the source mentions more than two variants, we select the two that are most different. Depending on the nature of the difference, variants were sometimes annotated using either phonological representations (e.g. PST = *ke* ~ *go*), schematic representations (e.g. NP = N Adj ~ Adj N), or combinations thereof (e.g. PRON.REFL = PRON-*mala* ~ PRON-*maŋ*). While we attempted to standardise these representations as much as possible, we found this to be a very challenging task, and one that will require further development if a database of this type were to be extended or reused. We also used a free-form Comments field to add additional description for most variables.

As described at the end of section §4.1, a substantial number of our grammatical variables contrast two forms that have differences in semantic range. Usually this involves one form being a semantic subset of the other, that is, a ‘neutralisation’ of some semantic distinction. We annotate such variables by representing the specific meaning, e.g. ‘NC:animate.human’, then representing the specific and neutralised forms schematically e.g. as ‘NC:animate.human ~ NC:animate’. This indicates that for the meaning/function of noun classification on human animates, one variant uses a specifically human classifier, while the other uses a more general animate classifier.

We also annotate the dialect groups in which the variation is exhibited. The first dialect (Region1) is associated with the first form (Var1), the second dialect (Region2) with the second form (Var2). We have not included details on these types of association in the database. The dialect groups are annotated as described by the author of the grammar and may denote named dialects, specific locations, or general regions. Where one group differs from all other groups, or an unspecified set of other groups, the latter is coded as ‘others’. Dialects may be associated with

each form in varying ways: the difference may be categorical (each dialect uses only their associated form) or a matter of different rates (one dialect uses a particular form more frequently than the other dialect).

E. Coding social distance

We coded social relations between dialect groups as either Close, Medium or Distant, as described in section §3.2. For most grammars, there was either just one major dialect relation to be coded (e.g. a single East/West distinction), or else a network of relations that appear to have Close social proximity (e.g. a cluster of clans that are densely inter-married). But a few reference grammars required multiple dimensions of dialectal proximity, where some groups appeared to be much closer together than others.

Information on the social relations between dialect groups was found in dialectology sections of introductory chapters, as mentioned above. The amount of information varied substantially from grammar to grammar, which led us to use the very coarse, three-point scale, as opposed to a more fine-grained or multidimensional index, which would be under-determined for many of our grammars. We estimated a single distance value (Close, Medium, Distant) by coding whatever information was available regarding several indicators of regular social interaction, such as intermarriage, mutual intelligibility of dialects, geographic distance, etc. These are listed in Table C1. For each indicator, if information was available we coded this on a scale of 0,1,2, with 0 representing the most regular social interaction. We then averaged the available indicators for each language to arrive at an overall index of social distance, divided into thirds and labelled as Close [0, 0.66]; Medium [0.67, 1.33]; Distant [1.33, 2].

Table C1. Indicators used to estimate social distance between dialect groups

INDICATOR	EXAMPLES OF CLOSE → DISTANT VALUES
Geographic distance	50km → 1000km
Mobility	Nomadic hunter-gatherers → Settled agriculturalists
Intermarriage	Exogamy → Endogamy
Economic relations	Extensive trade → Independent subsistence
Mutual intelligibility	Fully intelligible, groups mix or switch between dialects → Non-intelligible, groups interact using a lingua franca

F. Consistency of coding

All three authors of the paper coded multiple reference grammars. In addition, research assistance was provided by linguistics students Ben Volchok and David Osgarby, who coded or partially coded two reference grammars each.

The second author coded every grammar used in this study, to ensure consistency across grammars. In addition, 28 of the 42 grammars were independently coded by either the first or third author, with the second author then reviewing and adding to the initial coding. For the first 13 grammars coded at the beginning of the study, the first round of coding was done purely by reading, and the second round using keyword searches.

We performed intercoder reliability testing on three grammars, which suggested that the number of variables detected was not consistently coded, but the identification of variable types, and their status as dialectal or otherwise, were quite consistently coded. The implications of this for our study are addressed in section §3.1. The grammars of Domari (Matras 2012) and !Xun (Heine & König 2015) were each coded twice using the keyword method, while the Urarina grammar (Olawsky 2006) was coded once by reading, then a second time by keyword searches. Combining all three grammars, the first round of coding found 88 variables and the second found 105, including 39 of the same variables identified in the first round. This indicates the difficulty of consistently enumerating variables, and of arriving at the same set of variables when coding a grammatical description.

It is especially difficult to consistently enumerate variables in sections where the grammar provides dense detail on variation (for instance, a complex set of variant affixes encompassing several case categories among several dialects). In such instances it is impractical to capture every single variant and dimension of variation, and our approach has been to code representative examples from such complex sets. This may tend to favour dialectal variables over non-dialectal variables, since the dialectal variables provide richer information and may be preferentially selected by the coder as a key example. This may result in a slight over-representation of dialectal variables in the database.

However, the *number* of variables drawn from each source does not form part of our findings, and is unlikely to affect our findings given the structure of our statistical test (see

section §5. What is crucial for our study is that structural types and (non-)dialectal status should be consistently identified. When two coders do annotate the same variable, it is important that they arrive at the same codes for structural type and dialectal status. Of the 39 variables which were identified by both coders, 38 instances agreed on the Type of the variable (form, order, omission, construction). For dialectal status, there was agreement in 37 instances. These figures suggest that the coding of structural type and dialectal status was sufficiently reliable to support our findings.

G. Alternative regression model without random slopes

As described in the main paper, our regression model encounters convergence issues. These are likely due to the inclusion of random slopes for language family by structural type, which causes convergence issues (Bates et al. 2015), due to some language families where certain structural types have only a single data point (e.g. Arawan has only a single order variable). The convergence warnings are resolved by running further iterations of the model.

To ensure that our main results in no way depend upon these random slopes, we here report on a similar model but without the random slopes. The results of this model, illustrated in Table G1, are very similar to the main model: estimates are almost identical, and none of the significance levels change.

Table G1. Model coefficients of a mixed-effects regression predicting dialectal status

PARAMETER	REFERENCE LEVELS FOR PARAMETER ESTIMATE	ESTIMATE AND 95% CI (IN LOG ODDS)	P-VALUE
Intercept	Form variables in close contact	-0.07 [-0.58, 0.44]	p = 0.80
Type: Omission	Close contact	-1.69 [-2.07, -1.30]	**p < 0.001**
Type: Order	Close contact	-1.73 [-2.25, -1.21]	**p < 0.001**
Social distance (linear)	Form variables	-0.22 [-0.73, 0.28]	p = 0.37
Social distance (quadratic) x Omission	-	0.60 [-0.12, 1.19]	p = 0.08
Social distance (linear) x Order	-	1.38 [0.50, 2.25]	**p < 0.01**

G. Grammar source references

Beaumont, Clive H. 1979. *The Tigak language of New Ireland*. Canberra: Pacific Linguistics.
Blackings, Mairi & Fabb, Nigel. 2003. *A Grammar of Ma'di*. Berlin: Mouton de Gruyter.
Broadwell, George Aaron. 2006. *A Choctaw reference grammar*. Lincoln, NE: University of Nebraska Press.
Carlson, Robert. 2011. *A Grammar of Supyire*. *A Grammar of Supyire*. De Gruyter Mouton.
Chen, Weirong. 2020. *A grammar of Southern Min: The Hui'an dialect*. Berlin: De Gruyter.
Childs, G. Tucker. 2011. *A Grammar of Kisi: A Southern Atlantic Language*. *A Grammar of Kisi*. De Gruyter Mouton. (doi:10.1515/9783110810882)
Coupe, Alexandre R. 2007. *A grammar of Mongsen Ao*. Berlin: De Gruyter.
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.
Dienst, Stefan. 2014. *A Grammar of Kulina*. *A Grammar of Kulina*. De Gruyter Mouton. (doi:10.1515/9783110341911)
Dixon, R.M.W. 1988. *A grammar of Boumaa Fijian*. Chicago: University of Chicago Press.
Döhler, Christian. 2018. *A grammar of Komnzo*. Berlin: Language Science Press. (doi:10.5281/zenodo.1477799) (http://langsci-press.org/catalog/book/212) (Accessed April 19, 2019.)
Donaldson, Tamsin. 1980. *Ngiyambaa: The language of the Wangaaybuwan*. Cambridge: Cambridge University Press.
Doornenbal, Marius. 2009. *A grammar of Bantawa*. Meteren: Netherlands Graduate School of Linguistics.
Enfield, Nick J. 2007. *A grammar of Lao*. Berlin: Mouton de Gruyter.
Epps, Patience. 2008. *A grammar of Hup*. Berlin: Mouton de Gruyter.
Evans, Nicholas. 2003. *Bininj Gun-Wok: a pan-dialectal grammar of Mayali, Kunwinjku and Kune*. Canberra: Pacific Linguistics.
Feist, Timothy. 2015. *A grammar of Skolt Saami*. Helsinki: Suomalais-Ugrilainen Seura.
Ford, Lysbeth. 2011. *A description of the Emmi language of the Northern Territory of Australia*. Munich: Lincom Europa.
Forker, Diana. 2020. *A grammar of Sanzhi Dargwa*. Berlin: Language Science Press.
Heine, Bernd & König, Christa. 2015. *The !Xun language: A dialect grammar of Northern Khoisan*. Köln: Rüdiger Köppe Verlag.
Hellwig, Birgit. 2019. *A grammar of Qaqet*. Berlin: Mouton de Gruyter.
Hosokawa, Komei. 2011. *The Yawuru language of West Kimberley: A meaning-based description*. Munich: Lincom Europa.
Hualde, José Ignacio & Urbina, Jon Ortiz de. 2003. *A grammar of Basque*. Berlin: Mouton de Gruyter.
LaPolla, Randy & Huang, Chenglong. 2003. *A grammar of Qiang*. Berlin: De Gruyter.
Liddicoat, Anthony. 1994. *A grammar of the Norman French of the Channel Islands: The dialects of Jersey and Sark*. Berlin: Mouton de Gruyter.
Martin, Jack B. 2011. *A grammar of Creek (Muskogee)*. Lincoln, NE: University of Nebraska Press.
Matras, Yaron. 2012. *A grammar of Domari*. Berlin: De Gruyter.
McKay, Graham. 2000. Ndjebbana. In Dixon, R.M.W. & Blake, Barry J. (eds.), *The handbook of Australian languages*, vol. 5, 155–356. Melbourne: Oxford University Press.
Morey, Stephen. 2010. *Turung: A variety of Singpho language spoken in Assam*. Canberra: Pacific Linguistics.
Nikolaeva, Irina. 2014. *A grammar of Tundra Nenets*. *A Grammar of Tundra Nenets*. De Gruyter Mouton. (https://www.degruyter.com/document/doi/10.1515/9783110320640/html) (Accessed July 7, 2021.)
Olawsky, Knut J. 2006. *A grammar of Urarina*. Berlin: Mouton de Gruyter.
Peterson, John. 2010. *A grammar of Kharia: A South Munda language*. Leiden: Brill.
Rice, Keren. 1989. *A grammar of Slave*. Berlin: Mouton de Gruyter.
Ring, Hiram. 2015. *A grammar of Pnar*. Singapore: NanyangTechnological University. (PhD thesis.)
Saeed, John. 1999. *Somali*. Amsterdam: John Benjamins.
Sharp, Janet. 2004. *Nyangumarta: A language of the Pilbara region of Western Australia*. Canberra: Pacific Linguistics.
Shimelman, Aviva. 2016. *A grammar of Yauyos Quechua* (Studies in Diversity Linguistics). Berlin: Language Science Press. (https://langsci-press.org/catalog/book/83)
Smeets, Ineke. 2008. *A Grammar of Mapuche*. *A Grammar of Mapuche*. De Gruyter Mouton.
Smith, Ian & Johnson, Steve. 2000. Kugu Nganhcara. In Dixon, R.M.W. & Blake, Barry J. (eds.), *The handbook of Australian languages*, vol. 5, 355–490. Melbourne: Oxford University Press.

Valentine, Randy. 2001. *Nishnaabemwin reference grammar*. Toronto: University of Toronto Press.
Zariquiey, Roberto. 2018. *A grammar of Kakataibo*. Berlin: Mouton de Gruyter.

H. Other references in this supplementary information

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

Dryer, Matthew S. 1989. Large Linguistic Areas and Language Sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language."* John Benjamins 13(2). 257–292. (doi:10.1075/sl.13.2.03dry)

Hammarström, Harald & Donohue, Mark. 2014. Some Principles on the Use of Macro-Areas in Typological Comparison. *Language Dynamics and Change*. Brill 4(1). 167–187. (doi:10.1163/22105832-00401001)

Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2022. *Glottolog 4.6*. Leipzig. (doi:10.5281/zenodo.6578297) (https://glottolog.org/ accessed 2022-11-01) (Accessed November 1, 2022.)

Miestamo, Matti & Bakker, Dik & Arppe, Antti. 2016. Sampling for variety. *Linguistic Typology*. De Gruyter Mouton 20(2). 233–296. (doi:10.1515/lingty-2016-0006)