Oceania, the Pacific Rim, and the Theory of Linguistic Areas

BALTHASAR BICKEL and JOHANNA NICHOLS
University of Leipzig[*] and University of California, Berkeley

1. Introduction

A linguistic area is "a geographical region in which neighboring languages belonging to different language families show a significant set of structural properties in common, where the commonalities in structure are due to historical contact between speakers of the languages, and where the shared structural properties are not found in languages immediately outside the area (ideally where these include languages belonging to the same families as those spoken inside the area)" (Enfield 2005:190). That is, a linguistic area is defined by a group of *variables* (henceforth we use this term rather than *features*, *properties*, etc.) each of which constitutes an isogloss demarcating the area. Some linguists seek variables that form an isogloss bundle (e.g. Campbell et al. 1986, Joseph 1983, 2001); others do not (e.g. Emeneau 1956, Masica 1976), but nonetheless implicitly assume that some core part of the area should ideally emerge as located inside of all the isoglosses. Some works seek isopleths rather than isoglosses (van der Auwera 1998) and rank languages for the number of areal features they share. All of these approaches assume what we will call *categoriality* in the distribution of the defining variables: some value of a variable is present inside the area and absent outside of it (that is, in the neighboring languages outside of it).

Variable-defined areas present various problems. First, there are no criteria for deciding which are the diagnostic variables. This problem has an empirical side: the linguist needs to determine which variables are more and less frequent worldwide, which ones are most and least likely to diffuse, to be inherited; etc. It also has a statistical side. Suppose the linguist sorts through 200 variables and finds five that appear to be area-defining. Is this a significant result, or could one expect to find five out of 200 shared variables for any random set of languages and any random set of variables? The isogloss-bundled areal features standardly accepted for the Balkan and Mesoamerican language areas are selected from the entirety of the sound system, inventory of morphological forms, and basic syntactic inventory, a total set of elements that must number at least 200 and appears to be open-ended in practice. Half a dozen out of 200, or even 100, surveyed variables could easily cooccur in some set of languages by chance if they were at all frequent;

only if they were quite rare would it be unexpected for the set of languages to all show the entire half dozen variables. Our impression is that the classic Balkan features (to be listed below) include a few variables of sufficiently low frequency to be of diagnostic value, while the Mesoamerican ones include some that occur in one-quarter or more of the world's languages (head-marked nominal possession, non-verb-final basic word order), and one could expect five such to turn up in a survey of 200 or even 100 languages.[1] This issue has not had the discussion it deserves in the areal literature.

Second, a language may be a recent immigrant to an area and its speakers wholly involved in areal behavior such as bilingualism and code switching, yet the areal variables have not yet affected that language; does the linguist then draw a discontinuous isogloss quarantining the new language, disregard that language, or lower the standards for density of attestation of the criterial variables in the area? An example is Turkish spoken in Bulgaria, a core part of the Balkan linguistic area, by speakers bilingual in Bulgarian and/or Romani, both core Balkan languages. Balkanists have traditionally emphasized categorical variables found in all and only Balkan languages, with continuous isoglosses defining a coherent geographical area, and Bulgarian Turkish presents obstacles to the approach.

Third, the variables that can be identified as defining an area may be a motley set that raises few fruitful typological questions and does not fully capture the linguistic spirit of the area. An example of this is the classic Balkanisms (Joseph 1983:1, 2001:21): (i) postposed definite article, (ii) variant preposed future tense marker derived from a verb of volition, (iii) clitic doubling for objects, (iv) noun case mergers (especially displacement of genitive by dative; in the extreme situation, complete or near-complete loss of noun cases); (v) mid central vowel, (vi) lack of infinitive (finite subordinate clauses where most European languages use infinitives). It is true that identifying categorical Balkanisms is difficult because, except for Turkish, the Balkan languages are all related (as Indo-European) and much of what they have in common is inherited and shared with non-Balkan sisters. That said, the fact remains that the classic Balkanisms do not do a very complete job of defining the shared grammar that makes for the notable intertranslatability of Balkan languages.

Fourth, variables exhibiting the requisite isoglossic behavior may have to be defined as an abstraction which is in itself unlikely to be able to diffuse: an example is non-verb-final word order, a Mesoamerican areal variable identified by Campbell, Kaufman, and Smith-Stark (1986).

All in all, the variable-defined approach is unlikely to be able to define large, old, or inactive areas or areas with significant linguistic immigration very satisfactorily. This is because such areas are most likely to have diffuse boundaries, to

---

[1] A full statistical assessment will need to look at the worldwide frequency of the variable, the number of languages in the area, and the number of languages outside of but adjacent to the area (an area-defining feature cannot occur in any of these neighbors, though it can occur elsewhere in the world), and determine the probability of finding, say, five such variables given up to 200 attempts (or, perhaps more accurately, an open-ended number of attempts).

have internal nonconformities, to be typologically embedded in larger units, and to have confounding local divergence from areal norms.

Our approach turns the usual procedure on its head and defines variables from areas rather than vice versa. We define an area based on a theory of population and language spread and on information from other disciplines; hypothesize that it is a linguistic area; and test the hypothesis by seeking statistically non-accidental signals. We call this approach Predictive Areality Theory (PAT).

2.   Predictive Areality Theory

Each typological variable has its own history of and potential for change and spread, and therefore has its own distinct distribution over the world's languages. What underlies the impression of areality is that some such distributions overlap in a non-accidental way. If they overlap non-accidentally, one plausible explanation is shared history, by which we mean (any kind of) contact-induced change and/or shared inheritance (whether reconstructed and known or unreconstructible and unknowable). Such an explanation is a PAT holding for the specific regional overlap of the observed distributions. For a PAT to work, it must be grounded in what we know about population history from archaeology, genetics, ecology, geography, economics, demography, etc. Under this approach, then, areality is not a property of languages (e.g. 'in the Balkan Area' vs. 'not in the Balkan Area') but only a property of variables and sets of variables. In other words, areality is not, as under classical approaches, a typological observation. On the contrary, it is a theoretical predictor variable predicting observable typological distributions. The more the theory's predictions are statistically supported in such a series of predicted variables, the more robust the theory is.
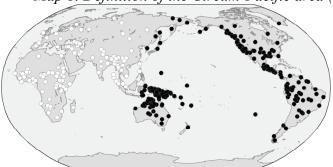
Regional overlap can be explained by a PAT only if we can demonstrate that the overlap does not result from (a) universal preferences (e.g. VP ~ PP order, or noun incorporation and head marking), (b) reconstructible shared genealogy, or (c) chance. We can use regular statistical inferencing to determine the probability of (c), but we need to control for (a) and (b). We control for (a) and (b) using standard typological methods: for (a) by rejecting typological variables as independent areal signals if they are known to be associated universally; for (b), i.e. for known genealogical relatedness, by constructing genealogically-balanced samples instead of random samples. The consequence of this sampling decision is that we cannot apply standard sampling theory and need to rely on randomization-based statistical methods. (See (Janssen et al. 2006) for further discussion.)

3.   The Pacific Rim as a Linguistic Area

In the 15 years since the first maps of numeral classifiers, head marking, and *n - m* personal pronouns were displayed to show a striking coast-hugging distribution all around the Pacific Rim (PR), a number of additional otherwise infrequent variables have been shown to have notably high concentrations in the Pacific-facing parts of the world. Yet the distributions of the variables that mark this putative area are manifestly not categorical or congruent. The area spans several

continents and lacks the compactness and centeredness of well-known smaller areas. Therefore, instead of attempting to trace area-defining isoglosses, we first define the area geographically and then ask whether any variables are significantly more (or less) frequent in the area than outside of it, and whether there are enough such to legitimately define an area. The rationale for grouping the entire Pacific Rim together as a single area includes human genetic and archaeological data indicating that the entire region was initially settled by migrations from ancient mainland Southeast Asia, continued to receive new colonizations from there up to and including the Austronesian expansion, and functioned as a contact and migration zone the whole time (Nichols 1997a, b, 2000, 2002).

We define the PR area as follows: Pacific-facing coast up to the lower slope of the far side of the major coast range (e.g. Andes, Sierras and Cascades, eastern Himalayas) or up to a coastal scarp (as in northern Australia). The Pacific Rim area is the more strictly coastal part of a larger area which we call the Circum-Pacific (CP) area. This comprises all of the Americas, Oceania (including Australia and New Guinea), and the mainland Asian Pacific Rim as just defined. That is, the CP area is the entire region anciently settled from coastal Southeast Asia and including the coastal Asian migration route. However, we exclude Southeast Asia (which we define as mainland Southeast Asia plus island Southeast Asia up to the Wallace line, i.e. including western Indonesia and the Philippines) from the CP area because it has considerably stronger historical and prehistorical ties to mainland Asia (Matisoff 1991, Enfield 2005) than to the other regions around the Pacific. We therefore expect Southeast Asia to pattern more often with Eurasia than with the CP. Drawing the boundary at the Wallace Line may appear arbitrary, but this is a natural breakpoint in our samples. Map 1 shows the definition of the CP area on a genealogically-balanced sample of languages.[2]

Map 1. Definition of the Circum-Pacific area (black dots) in our sample

There are five issues about this area (and similarly large areas) that now arise:

(a) Variance. Languages with PR or CP features everywhere coexist with languages lacking them. Classical definitions of areality (Masica 1976, Campbell

---

[2] The underlying table with genealogy and geography coding is available on our project website: http://www.spw.uzh.ch/autotyp/. All other codings discussed below are also deposited there.

et al. 1986, Joseph 2001; survey: Enfield 2005:190) assume near-100% consistency in variables across an area, but in reality within-area variance in otherwise good areal features is common. A clear example of such a variable is multiple possessive classes (more than one "inalienable" class of nouns; Nichols and Bickel 2005, POSSCL in the Appendix below). In fact, in the PR and CP areas, variance is expected and likely to have been an ancient and stable characteristic because the territory is almost entirely residual zone in the terms of Nichols 1992, and because the expansion of languages bearing PR features involved movement into already inhabited lands so that languages with PR features did not displace others but intermingled with them. Given this, we maintain that our areality prediction is confirmed by any statistically significant difference in frequencies inside vs. outside the area – regardless of variance inside the area.[3]

(b) Leakage. In certain places, PR variables "escape" into the nearby (and not-so-nearby) interior: syntactic noun incorporation (Houser and Toosarvandani 2006) in North America; ergativity [COMALN5], inclusive/exclusive pronouns (ExInDist, Bickel and Nichols 2005b) and reduplicated plurals in Australia; many variables in South America (where "PR" is a misnomer as there is almost never a discernible coastal cluster of PR variables). Under a PAT approach, this is expected because it has clear historical motivations. Wherever a spread zone abuts the PR zone (North America, Australia, inner Eurasia), "escaped" features are likely to spread far. Thus, for example, the spread of domestication from Mesoamerica impelled PR features eastward via the Caribbean coast. In our statistical survey below, we use the larger CP area as a predictor in order to capture at least the leakages on the American side.

(c) Greater variance and general diffuseness of PR variables in Oceania. A number of PR variables form notably denser clusters in the Americas than in Oceania, raising questions about the unity of the area and its specific history. Examples include high inflectional synthesis of the verb (Bickel and Nichols 2005a, SYN) and *n-m* personal pronouns (Nichols and Peterson 1996, 2005; NICNMP2). Rather than a problem, under a PAT approach this is again an expected phenomenon: Oceania has been inhabited longer than the Americas and domestication occurred earlier there than in the Americas (Denham et al. 2003), so the land was already linguistically and demographically saturated when the PR expansion began. In saturated conditions, new linguistic features had less impact and took root less readily.

(d) A troubling historical question: How could PR variables persist so long in an area when there are many cases of their loss within historically reconstructed language families that are younger than the PR? Rather than a shortcoming we see this as a defining property of diagnostic areal features: they are more persistent in areas than in families. This must be because their retention can be favored by

---

[3] Still, it might be useful to distinguish these general kinds of areal signals from signals that show strong within-area homogeneity (as measured for example by chi-square deviations from expected distributions within the area).

areal pressure, and because in linguistic areas they are prone to be transmitted not only by inheritance but also by substratal retention and diffusion.

4.   Survey

We tested our predictions about CP areality against the dataset available in the World Atlas of Linguistic Structures (WALS; Haspelmath et al. 2005), amended by our own richer datasets for the variables that we contributed ourselves to the Atlas. The WALS dataset is not (and is not meant to be) a genealogically balanced sample. Therefore we constructed an all-purpose sample for WALS, called 'WALSG', with one representative per genus (as that is defined in the Atlas). When there was a choice we opted for the language that is coded for the largest number of coded variables. For our own chapters, we used our standard genealogical sample in AUTOTYP, called 'GEN'. WALSG contains 193 languages, GEN 316. Using GIS software, we coded each language in both samples as belonging or not to the CP area. We used the larger CP area rather than its PR subpart because of the issue of leakage discussed above.

On an all-purpose sample, variables end up with many missing values. Of all variables available in WALS (or our versions of them) we selected those that have at least 150 (i.e. about 75%) non-empty values. This yields 75 variables.

The values of a typological variable can generally be lumped or split in various ways. For example, the variable of case alignment in (Comrie 2005) distinguishes marked from unmarked nominative/accusative alignment, while for different purposes one could treat them as the same and put them in opposition to several other alignments. In technical terms, these are all different ontologies derived from a single variable. In universals research we generally know which ontology is of interest to the prediction (e.g. accusative vs. other non-neutral alignment for predictions about which alignment type is preferred in agreement as opposed to case systems), but in areal typology we cannot know a priori which ontology will show areal overlap in its distribution. Re-ontologizing, or recoding, is of course only possible for multinomial variables and not all possible recodes are linguistically meaningful. With these constraints in mind, we recoded 23 of the 75 variables, with the number of recoded variants of each variable ranging from 2 to 6 (mode = 2). This yielded a total set of 100 variables. Note that some recodes increase again the number of missing values, but now these are logically necessary and not sampling gaps: for example, a binary recode of subtypes of accusative marking will have missing values only in languages that do not have accusative case alignment, but this is a fact of life and not a sampling problem.

We then tested our areality prediction against the 100 variables. That is, we surveyed not a hand-picked number of variables and not an open-ended set, but all variables available in testably high frequencies in both databases under genealogical sampling. For each variable, we tested whether there was a statistically significant difference between its frequencies in the Circum-Pacific and the rest of the world (i.e. Africa and non-Pacific Eurasia). For binary typological variables we used a 2x2 (typological variable x CP) Fisher Exact Test; for multinomial and

scalar variables we ran randomization-based chi-square and one-way anova tests, respectively, as described in (Janssen et al. 2006). We report the results in the Appendix, ranked by p-values.

5.   Results

When interpreting the results, we need to control for the fact that some variables might be universally correlated. We have not tested all possible universal correlations among the 100 variables, but the following word-order variables are well-known to correlate: DRYOBV0 ~ DRYGEN0 ~ DRYSOV0 ~ DRYSBV0 ~ DRY-ADP0 ~ DRYCOQ0 ~ DRYPQP01 ~ DRYPQP02 ~ VFIN ~ VFIN2 ~ VINIT ~ VINIT2; CORSEX01 ~ SIEGEN2 and SIEAPV2 ~ SIEVPA02 ~ POLYAGR are respectively the same or very similar variables coded by different researchers (see Appendix for what these labels stand for). What other correlations exist is an open question, one that needs extensive analysis. For now, we assume that 86 of the variables tested are distributionally independent of each other.

Running the same test on various recodings of the same variable increases the risk of familywise error of rejecting true null hypotheses. We controlled for this by applying Holm corrections to the *p*-values of each set of mutual recodings of a single variable (e.g., we corrected the *p*-values of all our 6 recodings of DRYSOV, Dryer's (2005) S-O-V order variable).

At a conventional .05 rejection level, we find that about 40% of the 86 variables that we assume to be independent show significant frequency difference between the CP area and the rest of the world. About 30% do so at a .01 level.[4]

6.   Conclusion

This has been an exercise in applying Predictive Areal Theory to a deep, old, and very large area which a priori presents many problems for areal analysis. We defined the PR and CP areas geographically, basing the definition and the geographical extent on what is known about human migrations and the settlement of the Pacific and the New World, then assembled a list of all variables which had enough data in an general-purpose database (WALS) and tested whether frequencies of variables in the area are significantly different from those outside the area. The outcome was that (depending on one's significance criterion) 30-40% of the variables yielded significance, and we regard each of these as a likely areal feature. This success rate is high enough to convince us that we have detected multiple symptoms of genuine areality. Note that the datasets were controlled for genealogical bias by an all-purpose sample, and this often meant that the actual dataset had to be shrunk, reducing the power of the statistical tests. It is possible that a sampling procedure that leads to larger samples would reveal more significant associations.

Our understanding is that the PR formed as coastally adapted people, and their

---

[4] Space limits make it impossible to include maps of the variables, but a sense of their actual distribution can be gained from the maps in WALS.

languages and cultures, spread out of Southeast Asia beginning late in the last glaciation and continuing into recent centuries with the Austronesian spread and the Chukchi spread to the Bering Strait. They spread coastally, as is shown by the striking coastal distributions of variables such as V-S order and multiple possessive classes. We tested for CP rather than more strictly for PR areality because leakage is such a pervasive problem as to obscure the linguistic boundary between the two (though not the geographical boundary, which we defined in advance).

All theories of areality take account of cultural, historical, and ecological factors as well as linguistic structure, but PAT differs in its crucial respects – defining areas geographically, no assumption of categoriality in variable distributions, testing all available variables for areality – because it was developed for work on large, old areas for which categoriality and neat isoglosses cannot be expected. Much work remains to be done, including development of statistical tools to define the minimum success rate that can be judged non-chance and to disentangle the PR from the CP. Even without these tools, however, the CP area has emerged as a clear linguistic area established by many independent variables.[5]

References

Bickel, Balthasar, and Nichols, Johanna. 2005a. Inflectional Synthesis of the Verb. In M. Haspelmath, M. Dryer, B. Comrie & D. Gil, eds., *The World Atlas of Language Structures*, 94-97.. Oxford: Oxford University Press.

Bickel, Balthasar, and Nichols, Johanna. 2005b. Inclusive/exclusive as Person vs. Number Categories Worldwide. In Elena Filimonova, ed., *Clusivity*, 47-70. Amsterdam: Benjamins.

Bickel, Balthasar, and Nichols, Johanna. 2002ff. *The Autotyp Research Program*. http://www.uni-leipzig.de/ ~autotyp/

Campbell, L., T. Kaufman, and T.C. Smith-Stark. 1986. Mesoamerica as a Linguistic Area. *Language* 62:530-570.

Denham, T. P. et al. 2003. Origins of Agriculture at Kuk Swamp in the Highlands of New Guinea. *Science* 301:189-193.

Dryer, M. S. 2005. Order of Subject, Object, and Verb. In M. Haspelmath, M. Dryer, B. Comrie & D. Gil, eds., *The World Atlas of Language Structures*, 330-34.

Emeneau, Murray B. 1956. India as a Linguistic Area. *Language* 32:3-16.

Enfield, Nicholas J. 2005. Areal Linguistics and Mainland Southeast Asia. *Annual Review of Anthropology* 34:181-206.

Haspelmath, M., M. Dryer, B. Comrie, and D. Gil, eds. 2005. *World Atlas of Language Structures*. Oxford: Oxford University Press.

Houser, Michael, and Maziar Toosarvandani. 2006. A Non-Syntactic Template for Syntactic Noun Incorporation. LSA Annual Meeting, Albuquerque.

---

[5] We thank Sven Siegmund and Anja Gampe for their help with the recoding of the WALS data.

Janssen, Dirk P., Balthasar Bickel, and Fernando Zúñiga. 2006. Randomization tests in language typology. *Linguistic Typology* 10. 419-440.

Joseph, Brian. 2001. Is Balkan Comparative Syntax Possible? In M. L. Rivero and A. Ralli, eds., *Comparative Syntax of Balkan Languages*. Oxford: Oxford University Press.

Joseph, Brian. 1983. *The Synchrony and Diachrony of the Balkan Infinitive*. Cambridge: Cambridge University Press.

Masica, Colin P. 1976. *Defining a Linguistic Area: South Asia*. Chicago: University of Chicago Press.

Matisoff, James A. 1991. Sino-Tibetan Linguistics: Present State and Future Prospects. *Annual Review of Anthropology*, 469-504.

Nichols, Johanna. 2003. Genetic and Typological Diversification of Language. In B. Joseph and R. Janda, eds., *Handbook of Historical Linguistics*, 283-310. London: Blackwell.

Nichols, Johanna. 2002. The First American Languages. *Memoirs of the California Academy of Sciences* 27:273-293.

Nichols, Johanna. 2000. Estimating Dates of Early American Colonization Events. In C. Renfrew, A. McMahon and L. Trask, eds., *Time Depth in Historical Linguistics, volume 2*, 643-663. Cambridge: McDonald Institute for Archaeological Research.

Nichols, Johanna. 1997a. Sprung from Two Common Sources: Sahul as a Linguistic Area. In P. McConvell and N. Evans, eds., *Archaeology and Linguistics: Aboriginal Australia in Global Perspective*, 135-168. Melbourne: Oxford University Press.

Nichols, Johanna. 1997b. Modeling Ancient Population Structures and Movement in Linguistics. *Annual Review of Anthropology* 26:359-384.

Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Nichols, Johanna. 1994. Ergativity and Linguistic Geography. *Australian Journal of Linguistics* 13:39-89.

Nichols, Johanna, and Bickel, Balthasar. 2005. Possessive Classification. In M. Haspelmath, M. Dryer, B. Comrie & D. Gil, eds., *The World Atlas of Language Structures*, 242-245.

Nichols, Johanna, and Peterson, David A. 1996. The Amerind Personal Pronouns. *Language* 72:336-371.

Nichols, Johanna, and Peterson, David A. 2005. Personal Pronouns: M-T and N-M Patterns. In M. Haspelmath, M. Dryer, B. Comrie & D. Gil, eds., *The World Atlas of Language Structures*, 544-551.

van der Auwera, Johan. 1998. Revisiting the Balkan and Mesoamerican Linguistic Areas. *Language Sciences* 20:259-27.

Balthasar Bickel and Johanna Nichols

Balthasar Bickel
Seminar für Allgemeine Sprachwissenschaft
Universität Zürich
Plattenstrasse 54
CH-8032 Zürich

balthasar.bickel@uzh.ch

Johanna Nichols
Department of Slavic Languages and Literatures, #2979
University of California
Berkeley, CA 94720-2979

johanna@berkeley.edu

Appendix: evidence for the CP area, ranked by corrected p-values

Variable	Values	Rough explanation (see WALS chapters for details)	Recode	WALS Chapter	Sample	N	CP evidence: corrected p	uncorrected
MADGAP	5	Missing common C		5	WALSG	175	9.97E-15	
MADVOI2	2	Voicing	1-2/3/4	4	WALSG	175	1.01E-14	5.07E-15
AUWEPI2	2	Epistemic modality verbal vs. affixal	1-2/3	75	WALSG	150	4.34E-11	
POLYAGR	2	Obligatory agreement with both A and P		22	GEN	276	4.08E-09	
POSSCL	2	Inflectional possessive classes		59	GEN	238	4.87E-08	
MADVOW	3	Size of vowel inventory		2	WALSG	175	2.96E-06	
DRYPOS0	3	Possessive prefix vs. suffix vs. both	1-2-3	57	WALSG	94	3.00E-06	
MADLAT2	2	Laterals		8	WALSG	175	8.22E-06	1.50E-06
SIEAPV2	2	Agreement with both A and P	1-2/3/4/5	104	WALSG	180	3.68E-06	4.11E-06
COMNUM5	5	Counting systems	1-2/3-4-5-6	131	WALSG	122	5.07E-05	1.84E-05
MADVOI0	2	Voicing in plosives vs. fricatives vs. both	2-3-4	4	WALSG	108	7.56E-05	7.56E-05
MADCON	scale	Number of consonants		1	WALSG	173	1.00E-04	
SYN	scale	Inflectional synthesis degree (w/o roles)		22	GEN	202	1.00E-04	
BAKADP2	2	Adpositions	1-2/3/4	48	WALSG	179	1.06E-04	3.54E-05
SIEPAS	2	Passive		107	WALSG	178	2.78E-04	
NICMTP2	2	m/t-pronouns	1-2/3	136	GEN	185	4.68E-04	
DRYGEN0	2	GenN order	1-2	86	WALSG	163	8.56E-04	
MIEASY	7	Asymmetry types in NEG		114	WALSG	170	9.69E-04	
MADLAT0	4	Lateral series	2-3-4-5	8	WALSG	141	1.05E-03	
COMALN5	5	Case alignment of nouns (ACC collapsed)	1-2/3-4-5-6	98	WALSG	164	1.88E-03	1.05E-03
SIEAL10	5	Alignment in agreement	1-2-3-4-5-6	100	WALSG	140	2.07E-03	
ExInDist	2	Incl/Excl-Distinction		*	GEN	289	2.14E-03	
HAAEVD2	2	Evidentials	1-2/3	78	WALSG	170	2.92E-03	
SIEZER2	2	S agreement	1-2/3/4/5/6	103	WALSG	180	3.24E-03	1.62E-03
SIEVPA01	2	Agreement	0-2/3/4/5	102	WALSG	180	3.86E-03	3.42E-03
SIEVPA02	5	A and/or P agreement	1-2-3-4-5	102	WALSG	140	3.86E-03	1.93E-03
CORASS01	2	Semantic vs. semantic and formal gender	2-3	32	WALSG	53	4.87E-03	2.86E-03
MADTON02	2	Tone	1-2/3	13	WALSG	169	5.72E-03	
DRYPRE0	3	Affix position trend	2/3	26	WALSG	145	6.49E-03	
DRYSOV0	6	S,V,O orders	1-2-3-4-5-6	81	WALSG	145	7.50E-03	1.25E-03
HAAEVC0	5	Evidential marking types	2-3-4-5-6	78	WALSG	78	7.73E-03	
DRYPOS2	2	Possessive affixes	1/2-3	57	WALSG	151	9.97E-03	9.97E-03

Code	n	Description	Values	#	Source	N	p	p2
ExAsPers	2	incl/excl as person	1-2/3/4	*	GEN	289	0.01025	
MADFRV2	2	Front rounded V	3/4/7-1/2/5/6	11	WALSG	174	0.01100	5.50E-03
VINIT2	2	V-initial or free order	1/2	81	WALSG	175	0.01320	2.64E-03
DRYSBV0	2	SV vs VS order	1/2	82	WALSG	172	0.01374	
DRYPRO0	5	Type of pronominal subject expression	1-2-3-4-5	101	WALSG	152	0.01406	
LocPOSSU2	2	Double-Marking possessor and object		23	GEN	248	0.01790	
DRYADP0	2	Adposition: post vs. pre vs. in	1-2-3	85	WALSG	164	0.02322	
SIEZER0	2	Nonzero vs. zero in 3sAGR	2-3/4/5/6	103	WALSG	135	0.02349	0.02349
ANDANG2	2	Velar nasal present	1/2-3	9	WALSG	168	0.02527	0.01263
IGGNUM0	2	Case	1-2/3/4/5/6/7/8	49	WALSG	172	0.02657	0.01328
WOFREE	2	Free word order	1/2/3/5/6-7	81	WALSG	175	0.03316	0.00829
DOBOPT	2	Inflectional Optatives		73	WALSG	157	0.03577	
BAECSY01	3	Case syncretism degree	2-3-4	28	WALSG	64	0.03918	0.01959
DRYOBV0	2	OV vs VO	1-2	116	WALSG	169	0.04322	
CORSEX01	2	Gender	1-2/3	31	WALSG	147	0.04749	0.01583
AUWHOR	4	Type of hortative system		72	WALSG	153	0.05141	
SIEGEN2	2	Gender	1/2/3/4/5-6	44	WALSG	178	0.06509	0.03255
CORSEX	3	no gender vs. sex-based vs. other		31	WALSG	147	0.07560	0.03780
VFIN2	2	V-final or free order	1/6/7-2/3/4/5	81	WALSG	175	0.07932	0.02644
CORNUM	scale	Number of genders		30	WALSG	143	0.08000	
DRYDEM	6	DemN orders		88	WALSG	174	0.09089	0.04545
DRYDEM0	2	Demonstrative initial vs. final	1/2-2/4	88	WALSG	163	0.09089	0.07309
NICNIMP2	2	n/m-pronouns	1-2/3	137	GEN	185	0.10616	
DRYCOQ0	2	WH initial	1-2	93	WALSG	140	0.11897	
MIAuDist	2	Minimal/augmented system		*	GEN	289	0.13819	
SIEGEN0	5	Gender across person and number categ.	1-2-3-4-5	44	WALSG	55	0.14064	0.14064
IGGNUM	scale	Number of cases		49	WALSG	172	0.14340	0.14340
DRYNPL	9	Coding type of plural		33	WALSG	164	0.16180	
COMALP0	5	Pronoun alignment (ACC collapsed)	1-2/3-4-5-6	99	WALSG	147	0.16335	
BAKADP02	3	Adposition agreement	2-3-4	48	WALSG	152	0.16556	0.08278
MADPRS0	6	Presence of uncommon consonants	2-3-4-5-6-7	19	WALSG	28	0.23562	
PREROLE	2	Some agreement prefixed		22	GEN	160	0.24503	0.25715
MADTON01	2	Simple vs. complex tone	2-3	13	WALSG	56	0.25715	0.25715
DANPLU04	3	Types of expressing plural an pronouns	3-4/5/6-7/8	35	WALSG	166	0.28073	0.05615
BAEPSY01	2	Subject agreement syncretism	2-3	29	WALSG	124	0.28713	0.14356
AUW1MP2	2	Morphological imperative	1/2/3/4-5	70	WALSG	170	0.37041	
LocU2	2	Double-Marking object		25	GEN	245	0.40627	

Code		Description	Recode	N	Source	WALS		
SONNON2	2	Nonperiphrastic causatives	1-2/3/4	111	WALSG	158	0.40891	
DRYNUM0	2	NumN vs. NNum	1-2	89	WALSG	161	0.43265	
MADSYL	3	Complexity of syllables		12	WALSG	167	0.45860	
DRYPQP01	2	Position of Q-particle	1-2-3-4-5	92	WALSG	88	0.46069	0.30851
DRYPQP02	4	Position of Q-particle ('early' collapsed)	1/3-2-4-5	92	WALSG	88	0.46069	0.23034
MADUVU2	2	Uvular C	1-2/3/4	6	WALSG	175	0.48796	0.24398
MADUVU0	3	Uvular C series	2-3-4	6	WALSG	35	0.48796	0.25946
BAKADP01	2	Agreement on adpositions	2-3/4	48	WALSG	152	0.49592	0.49592
VINIT	2	V-initial	3/4-1/2/5/6/7	81	WALSG	175	0.53143	0.26571
DRYTAA2	2	Tense/aspect inflection	1/2/3/4-5	69	WALSG	176	0.53531	
MADGLO0	2	Glottalized C	1-2/3/4/5/6/7/8	7	WALSG	175	0.59847	
BAEPSY02	2	Subject agreement	1-2/3	29	WALSG	171	0.60374	0.60374
LocPOSS2	2	Double-Marking possessor		24	GEN	244	0.61984	
DRYCAS	9	Morphological type of case		51	WALSG	165	0.62349	
HAJNAS	2	Nasal vowels		10	WALSG	149	0.69216	
CORSEX02	2	Sex-based vs. non-sex-based gender	2-3	31	WALSG	53	0.74044	0.74044
SIEAPV0	2	A before P vs. P before A in agreement	2-3	104	WALSG	72	0.79550	0.79550
AUWPRH22	2	Dedicated prohibitive	1-2/3/4	71	WALSG	154	0.85358	0.42679
AUWPRH21	2	Prohibitive as imperative	1/2-3/4	71	WALSG	154	0.85358	0.51540
DRYADJ0	2	AdjN vs Nadj	1-2	87	WALSG	161	0.87035	
VFIN	2	V-final order	1/6-2/3/4/5/7	81	WALSG	175	0.87767	0.87767
MIESYM	3	Symmetric vs. asymm. vs. mixed negation		113	WALSG	170	0.91513	
DANPLU01	7	Type of plural coding on subject pronouns	3-4-5-6-7-8	35	WALSG	166	0.98075	0.24519
DRYPOQ2	2	Interrogative - declarative distinction	1/2/3/4/5/6-7	116	WALSG	155	1.00000	
ANDANG0	2	Velar nasal banned from initial position	1-2	9	WALSG	79	1.00000	
BAECSY02	2	Case syncretism	2/3-4	28	WALSG	64	1.00000	1.00000
DANPLU02	2	Subject pronouns (present or not)	1-2/3/4/5/6/7/8	35	WALSG	175	1.00000	1.00000
DANPLU03	2	Person-number vs. person stems	4/5/6-7/8	35	WALSG	146	1.00000	0.50910
DANPLU05	3	Person and number coexponence in pron.	3/4-5/7-6/8	35	WALSG	166	1.00000	0.66323
MADFRV0	3	Type of front rounded vowel	2-3-4	11	WALSG	9	1.00000	0.90139
DRYNEG2	2	Single vs. double negation	1/2/3/4/5-6	112	WALSG	166	1.00000	1.00000

Explanations:

'Recode': the definition of how values in the WALS database were recoded. The values are shown by the numerical labels they have in WALS and '/' means that values were collapsed whereas '-' means they were kept distinct; values that were excluded are those that are not listed here.

'N': the number of languages with a non-missing value for the variable in the sample.

* Bickel and Nichols (2005b), corresponding to WALS Chapters 39 and 40 by Michael Cysouw