New Perspectives on the Peopling of the Americas, ed. by Katerina Harvati, Gerhard Jäger,
Hugo Reyes-Centeno. Words, Bones, Genes, Tools: DFG Center for Advanced Studies Series.
© 2018, Kerns Verlag, Tübingen, ISBN: 978-3-935751-28-5.

CHAPTER 1

Morphological complexity of languages reflects the settlement history of the Americas

Johanna Nichols[a] *and Christian Bentz*[b]

Abstract

Morphological complexity is widely believed to increase with sociolinguistic isolation, and to decrease with language spreads and absorption of L2 adult learner populations. However, this can be assessed only for communities with well-described histories. Morphological complexity has also been shown to be greater in higher-altitude languages, which are often sociolinguistically isolated, so we use altitude as an empirically determinable proxy for sociolinguistics. In past research, only a very few small locations have been surveyed and the measures of complexity used were family-specific and not easily generalizable. We apply several improved measures of complexity and show that the correlation holds, especially in the Andean regions of South America. We discuss the implications for the South American pattern for the settlement of the Americas and post-settlement prehistoric population formation.

Resumen

Se cree que la complejidad morfológica incrementa en función del aislamiento sociolingüístico y disminuye con la propagación del lenguaje al igual que con la absorción de poblaciones con aprendices adultos de un segundo idioma. Sin embargo, esta suposición solo se puede evaluar en poblaciones con historias bien descritas. También se ha demostrado que la complejidad morfológica es mayor en lenguajes de altitudes más elevadas, mismos que están frecuentemente aislados sociolingüísticamente, por lo que en este estudio usamos la altitud geográfica como un proxy empíricamente determinable en la sociolingüística. En investigaciones previas solo se han estudiado algunas localidades pequeñas y las medidas de complejidad que se emplearon fueron específicas para cada familia y no fácilmente generalizables. Nosotros aplicamos una serie de medidas de complejidad mejoradas, y mostramos que la correlación se mantiene, especialmente en las regiones andinas de Sudamérica. Así mismo, discutimos las implicaciones que esto tiene en cuanto al patrón sudamericano para el poblamiento de América, así como para la formación de poblaciones prehistóricas después del asentamiento.

[a] University of California, Berkeley, USA.
[b] University of Tübingen, Germany.

INTRODUCTION

Languages vary considerably in their morphological complexity, and mean complexity levels vary considerably from language family to language family, area to area, and continent to continent. The causes of this variability include the sociolinguistics of contact vs. isolation, constraints on L2 vs. L1 learning, a certain range of random variability, and the inherent conservatism of language transmission, which preserves inherited patterns by default. Due to this last factor historical linguists assume that significant differences in complexity—especially where they characterize not single languages but whole populations of languages—take time to develop. This makes it possible to use modern distributions of complexity to infer aspects of linguistic prehistory.

Here we present a pilot study applying three different complexity measures in order to map out the distribution of morphological complexities across South America (for a fuller typology of complexity measures see Sinnemäki 2011: 23, Karlsson et al. 2008). We then relate the variance in complexities to the geographic dimensions of altitude and longitude via Pearson correlation analyses. This is a first assessment of interesting patterns, and a first step towards more fine-grained and elaborate statistical modeling. We further discuss implications of our preliminary results for the settlement of the Americas.

MORPHOLOGICAL COMPLEXITY MEASURES

Inventory complexity

The simplest measure of complexity is *inventory complexity*, a.k.a. *taxonomic complexity*, which is the total number of units in a system, for some subsystem or set of subsystems of a language. For instance, for phonology, this could be the number of phonemes, or the number of consonants, or the number of tones, etc. For more detailed discussions of inventory complexity, see Sinnemäki (2008) and Nichols (2009). Here we use two measures of inventory complexity: one extracted from the Autotyp database (Bickel and Nichols 2002 ff.), and one from the World Atlas of Language Structures (WALS, Dryer and Haspelmath 2013).

From Autotyp we surveyed six morphological variables: inflectional synthesis of the verb (represented in both Autotyp and WALS), presence of plural marking on nouns, presence of dual marking on nouns, presence of numeral classifiers, presence of gender agreement, presence of autogender on nouns. The morphological complexity measure is then the sum of the individual values. We are here mainly interested in morphological complexity. However, we additionally surveyed four phonological variables: number of contrastive consonant series, size of vowel inventory (S/M/L), number of contrastive tones, complexity of syllable structure (represented in both Autotyp and WALS); as well as two syntactic variables: number of default alignments of A and S, and number of basic word orders. The additional variables allow us to further assess whether the morphological complexity trends also hold beyond morphology. The whole set of variables is that used in Nichols (2009) and Nichols (2015).

From WALS we surveyed 28 features exclusively relevant to morphology. These include ordinal features such as "Number of Genders" (Chapter 30A) or "Number of Cases" (Chapter 49A), which run from 0 to "5 or more," and 0 to "10 or more" respectively, and also binary features such as "The Future Tense" (Chapter 67A), which merely give "absent" or "present," i.e., 0 or 1. To make different variables comparable, we normalize values to the interval [0,1]. The overall morphological complexity score per language is then the average across the (available) features. For details on all the features used and the methods see Bentz et al. (2016).

Similar inventory complexity measures of both the Autotyp and WALS have been used before in typological studies, for instance, by Nichols (2009), Lupyan and Dale (2010), and Bentz and Winter (2013).

Opacity or non-transparency

A closer approximation to complexity as it is encountered by language learners is *opacity*, or *non-transparency*, which can be measured as the number of departures from the ideal mapping of one form ↔ one function, or one form ↔ one meaning. For instance, to mark the plural of nouns English uses two major strategies, the productive *-s* plural as in *cats* or *dogs* and the unproductive vowel change as in *foot : feet*; and a few minor strategies such as the *-en* of *oxen* and *children* or the zero marking in *sheep*, *deer*, and a few others. Hausa, on the other hand, has 20 distinct kinds of plural marking (Newman 2000; Caron 2013), which are mostly non-predictable and must be lexically stipulated for each noun—a much less transparent system. Another example is gender categories and their markers. Avar (Nakh-Daghestanian: eastern Caucasus) has three genders, all semantically predictable (human males, human females, and non-human nouns) and formally regular (the markers are *w*, *j*, and *b* respectively). This is a perfectly transparent gender system; the only non-transparency is the presence of gender at all, since gender agreement marks no function or category other than itself and is basically unnecessary in language (Corbett 1991). In contrast, the distant sister language Tsakhur has four genders, two of them semantically predictable and two of them arbitrary, and each gender marker has two different forms used in different morphological contexts; in addition, gender marking is prefixal for some verbs and infixal for others (Dobrushina 1999). This opacity related measure of morphological complexity is further described in Appendix 1 (end of chapter) and in Nichols (2015).

Word entropy

The third measure used here is *word entropy*, more precisely the entropy of *unigrams*, i.e., runs of alphanumeric UTF8 characters delimited by white spaces. These are extracted from three parallel corpora: 1) the European Parallel Corpus (EPC, Koehn 2005); 2) the Universal Declaration of Human Rights (UDHR); and 3) the Parallel Bible Corpus (PBC, Mayer and Cysouw 2014). The sample we use here includes 220 texts of 162 South American languages (ISO 639-3 codes). The unigram word entropy of a given text *T* (indirectly representing a language) is then calculated as

$$H(T) = -\sum_{i=1}^{V} p\,(w_i)\; log_2\, p\,(w_i),$$

where *V* is the size of the vocabulary of word types (unique unigrams), and $p(w_i)$ is the probability of a given word type *i*. *H(T)* reflects the unpredictability associated with words in written language production. If a language has a wide range of different word types with low frequencies of occurrence (low probabilities) then *H(T)* is high. If a language has few different word types with high frequencies then *H(T)* is low. See Bentz et al. (2017) for more detailed explanations and mathematical formulations. The bottom line is that a language with high morphological complexity, i.e., very productive morphological marking strategies, has a wide range of word forms, higher word unpredictability, and hence higher entropy—as a general trend. For instance, in English there is only one possible morphological modification of the noun *tree*—namely *trees* to mark plural (potentially also the clitic *'s* to mark possession). In German, on the other hand, the noun *Baum* can be modified to *Baum(e)*, *Baum(e)s*, *Bäume*, *Bäumen*. As a consequence, German has a wider range of word forms and higher word entropy. This relationship between word entropy and morphological complexity is further illustrated in Bentz et al. (2016). Word entropy thus reflects word form complexity in actual written language production. For more details see also Bentz et al. (2015), Bentz and Alikaniotis (2016), Bentz et al. (2016), Bentz et al. (2017), and Koplenig et al. (2017).

COMPLEXITY AND ALTITUDE

Morphological complexity has a number of interesting distributional correlations with geography. It is higher in the Americas than in the Old World (Nichols 2009; Donohue and Nichols 2011), it increases when measured with unigram entropy with higher latitude worldwide (Bentz 2016), it increases with altitude (Nichols 2013, 2016), and it forms a worldwide west-to-east cline in the northern high latitudes. These last two patterns are described here with a focus on the Americas, while taking the Caucasus region as a testbed explored in earlier studies.

The Caucasus as a testbed

Where mountain highlands host permanent settlements, complexity levels tend to be higher in the highlands than in the surrounding lowlands.

In Daghestan (the eastern part of the Caucasus), morphological non-transparency is highest in the highlands (Nichols 2013, 2016). Figure 1 shows gender classification and gender agreement categories in Avar, Tsakhur (discussed above), and their sister languages, their locations, and their complexity levels. The important lowland contact languages—Avar, Andi and its closest sisters, Lezgi, and Udi (which is now a small enclave language, known for its exotic feature of endoclisis (Harris 2002), but in the early to mid first millennium it appears to have been an inscriptional and inter-ethnic language of some importance)—have the lowest non-transparency levels, and Tsakhur and others in the highlands have the highest. Figure 2 plots non-transparency of gender marking against altitude for the languages of Daghestan. There is a moderate correlation on the verge of significance (r = 0.36, p = 0.06): languages spoken at higher altitudes tend to be less transparent. Note that this carefully collected lan-

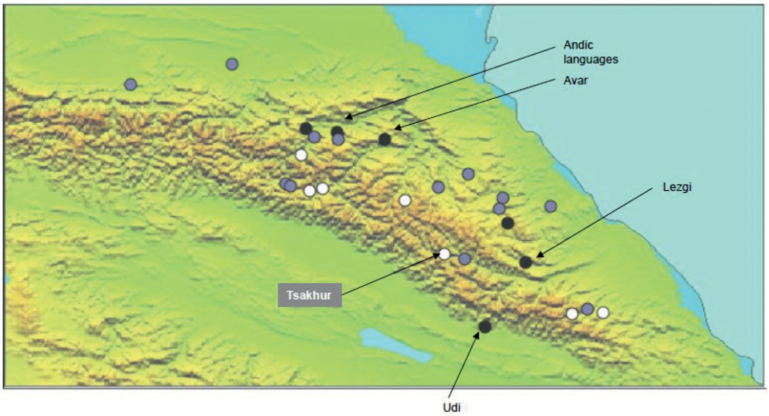


Fig. 1.
Topographic map of the eastern Caucasus, showing Nakh-Daghestanian languages. Black = most transparent, gray = intermediate, white = least transparent. Language labels indicate major contact languages, and Tsakhur, discussed above in text.

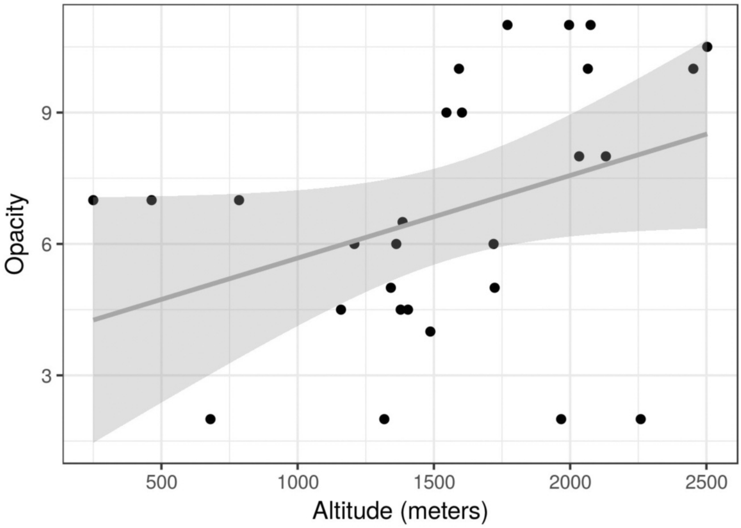


Fig. 2.
Non-transparency and altitude in the languages of Daghestan (eastern Caucasus). N = 28. Adopted and modified from Nichols (2013, 2016). The Pearson correlation is r = 0.36 (p = 0.06). Gray line indicates a linear regression model with 95% confidence band.

Table 1.
Opacity levels of noun, pronoun, and verb inflectional paradigms in languages along the Andi Koisu (left bank) and Avar Koisu rivers in Daghestan. Languages are ordered by relative position along rivers (headwaters at the top). Branches of Nakh-Daghestanian are in parentheses.

	Andi Koisu	**Avar Koisu**	**Opacity**
Highland:		Hunzib (Tsezic)	32
	Hinuq (Tsezic)		24
	Tindi (Andic)		15
	Godoberi (Andic)		14
	Karata (Andic)		16
Lowland:	Avar (standard) (Avar)		18

guage sample is necessarily small (N = 28), which certainly impedes statistical significance.

In the eastern Caucasus, towns and villages are strung out along the major river canyons, and in the highlands every small town or village often has its own unique language. Table 1 shows languages along the Andi Koisu and Avar Koisu for which morphological opacity has been calculated. Similar vertical chains can be traced along the right bank of the Avar Koisu and along the Samur in the south. These involve fewer languages than the plot shown in Figure 1, but the relative opacity levels and altitudes are directly comparable here, and they show the same sort of correlation, with higher opacity at higher altitudes.

Complexity levels in these examples correlate with altitude, but altitude itself is not the causal factor. Rather, it is the greater sociolinguistic isolation of highland villages that accounts for the greater complexity of their languages (Nichols 2013, 2016). These villages—difficult of access, distant from markets, with short growing seasons, and almost entirely endogamous—receive almost no immigrants and therefore almost no L2 learners, and it is intake of L2 learners that most clearly tends to decomplexify languages (Dahl 2004; McWhorter 2007, 2016; Lupyan and Dale 2010; Trudgill 2011; Bentz and Winter 2013; Bentz et al. 2015; Bentz and Berdicevskis 2016).

Altitude and complexity in South America

Turning to South America, morphological complexity correlates with altitude as well. Figure 3 illustrates relationships between altitude and three morphological complexity measures. The first is word entropy; the other two are inventory complexities based on Autotyp and WALS (opacity metrics for these are not available at the time of writing).

The correlation for altitude and word entropy is moderate and significant (r = 0.36, p < 0.0001), while weaker and non-significant for inventory complexity based on Autotyp (r = 0.19, p > 0.05), and based on WALS (r = 0.30, p = 0.17). Again, the last two run into the problem of data sparsity. While word entropy can be calculated for 162 South American languages in this sample, Autotyp and WALS only give 31 and 40

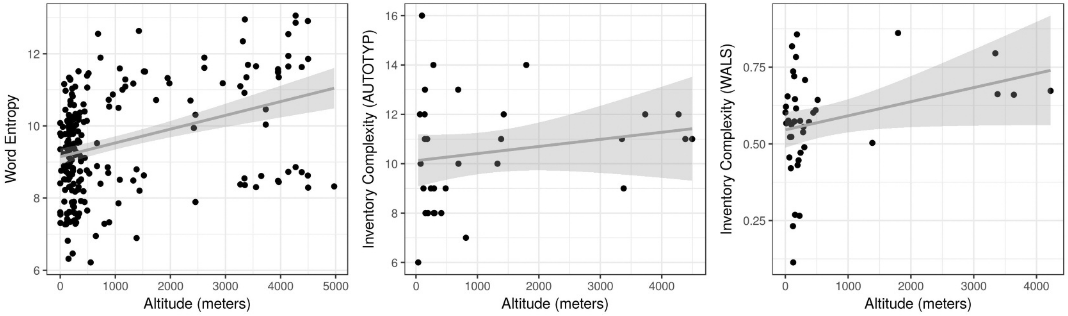


Fig. 3.
Correlations between altitude (x-axis) and scaled morphological complexity measures (y-axis) in South America. The three measures are word entropy for 162 languages (left panel), inventory complexity based on the Autotyp database for 31 languages (middle panel), and inventory complexity based on the World Atlas of Language Structures (WALS) for 40 languages. The Pearson correlations are from left to right: r = 0.36 ($p < 0.0001$), r = 0.19 ($p > 0.05$), r = 0.30 ($p = 0.17$). We here use the Bonferroni method to control for multiple testing, i.e., the original p-value of the Pearson correlation is multiplied by the number of tests (3 in this case). Gray lines indicate linear regression models with 95% confidence bands.

data points respectively. This certainly explains the drop in significance. Importantly, all three measures agree in their positive correlation of morphological complexity and altitude.

Figure 4 visually illustrates the altitude effect for South America. It depicts the 162 South American languages of the word entropy sample on a three dimensional topographic map. Larger dots correspond to higher word entropy, i.e., complexity. The altitude/complexity relationship is driven by languages spoken in the Andean region, which are generally morphologically complex. Languages of the Amazonian areas in the lowlands are, on average, less morphologically complex.

Zooming into the Andean highlands, we further find that there can be subtle differences between language groups in terms of the inventory and opacity metrics. Table 2 shows succession to power and non-transparency levels in the southern Andean highlands.

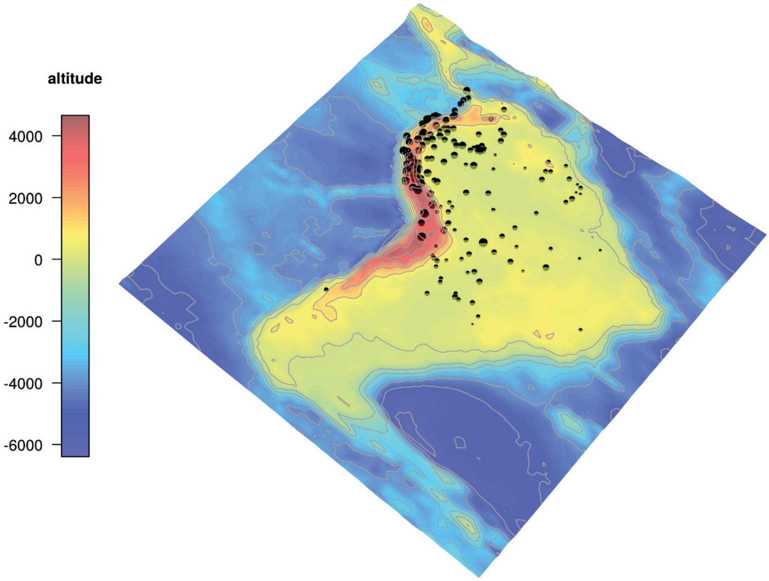


Fig. 4.
Topological plot of altitude and word entropy across South America. Every black dot corresponds to one of 162 languages in the word entropy sample. Larger dots correspond to higher entropy, as a proxy for higher morphological complexity.

Table 2.
Language succession and decomplexification in the Andean altiplano. Overall: combination of phonological, morphological, and syntactic inventory sizes based on Autotyp; Morphological: morphological inventory size only.

Language, economy, history	Overall	Morphological
Uru-Chipaya: Chipaya. Riverine/lacustrine.	22	12
Non-state language. Long present pre-contact.		
Aymaran. Agricultural		
Jaqaru: Non-state. Long indigenous.	22	14
Aymara: Pre-Inka state language (c. 500–1000 CE)	20	12
Quechuan. Agricultural		
Peripheral: Long indigenous (Huallaga, Tarma)	18	12
Central: Coastal trade variety spreads as interethnic language (Cochabamba, Ayacucho, Cuzco)	18-19	10-11
Central: Official language of Inka empire (13th–16th centuries) (Imbabura)	16	9

The Uru-Chipaya family, well entrenched in the highlands for at least the last few millennia and probably longer, is sociolinguistically isolated and quite complex. Equally complex is Jaqaru of the Aymaran family, long indigenous in the highlands. Its sister Aymara, the pre-Inka state language, is less complex. Peripheral Quechuan languages and the coastal trade language that spread after the pre-Inka state collapse are less complex, and the central varieties descended from the Inka imperial language still less complex. For the linguistic history of the Andes, see Adelaar and Muysken (2004). Complexity levels among neighboring languages range from almost as high as Chipaya in the south (Mapudungun, isolate of Patagonia; Matacoan and Guaycuruan families, northwestern Argentina) to much lower in the north (Arawakan, Panoan, and Jivaroan families, all of Peru; Barbacoan family, Colombia). Overall, then, it seems that morphological complexity started out high in the Andean *altiplano* and decreased with contact and especially statehood; complexity levels were slightly lower to start with in southern neighbors and much lower to start with in the high-contact region of Peruvian upper Amazonia. On the other hand, Van Gijn et al. (n.d.) give some evidence that contact in upper Amazonia might have led to diffusion of morphological marking strategies rather than systematic reduction. In very different ways, imperial Quechua and the diverse languages of upper Amazonia are high-contact languages, with immigrants and language learners absorbed by state expansion in the highlands and traditional intermarriage and mobility in Amazonia.

Some hedges on Table 2 are in order. First, the table and the brief descriptions of economy and history are hypothesis-raising only; close ethnographically and politically/economically based description and comparison will be required before statistically sound comparison figures can be drawn up. In any event, such factors as statehood, economic reach, etc. are ultimately proxies for the causal sociolinguistic factors of inter-ethnic communicative function of the language and absorption of adult second language learners by the community, which are difficult to ascertain from historical and (especially) archaeological evidence. Second, the strongest support comes from comparison within families; we assume that as a community gains higher-level political organization and/or economic power its language becomes morphologically less complex compared to its ancestral state, and sister languages with isolated societies may still reflect that state. Third, comparison here is intended only within the highland Andean area and very nearby. Mean absolute complexity levels for these languages are high relative to those for, e.g., South America as a whole, as shown in the preceding paragraphs; the point here is that within that area evolution is in the direction of less complexity for languages that have served as vehicles of inter-ethnic communication. As an example of these points consider Jaqaru and Aymara in the table. The two languages are closely related, closely enough that their separation may not antedate the beginnings of statehood for their speakers. But, whatever the causes of their different locations and histories, Aymara was a state language for some 500 years and Jaqaru was not, and it is this sociolinguistic status that the complexity levels are claimed to reflect.

COMPLEXITY AND LONGITUDE

Figure 5 shows overall levels of inventory complexity (Autotyp-based) for 193 languages surveyed worldwide. In the high latitudes (above 40° N) there is a fairly strong correlation of complexity and longitude ($r = 0.42$, $p < 0.001$): complexity levels are low in western Europe and increase to high in eastern North America. The mid and low latitudes of the northern hemisphere show a very similar correlation ($r = 0.42$, $p < 0.01$), while for the southern continents (Africa, Australia-New Guinea-Oceania, South America) this is considerably reduced and non-significant ($r = 0.19$, $p = 0.25$). Note that this reduction and non-significance is not due to smaller sample size ($N = 82$, compared to $N = 65$, and $N = 45$).

The longitude/complexity relationship is then a mainly northern phenomenon, and it links the entire high-latitude northern hemisphere in a single typological trend. The cline obtains within Eurasia, as seems to be in line with the deep connections posited by Jäger (2015) based on lexical similarity, but continues beyond into North America. However, we would connect differences in complexity to sociolinguistics more than to common descent.

Very similar correlations between longitude and language structure are exhibited by all other morphological phenomena allowing fine or multivariate breakdowns that we have surveyed: inflectional person (very strong correlation), causativization as preferred realization of the

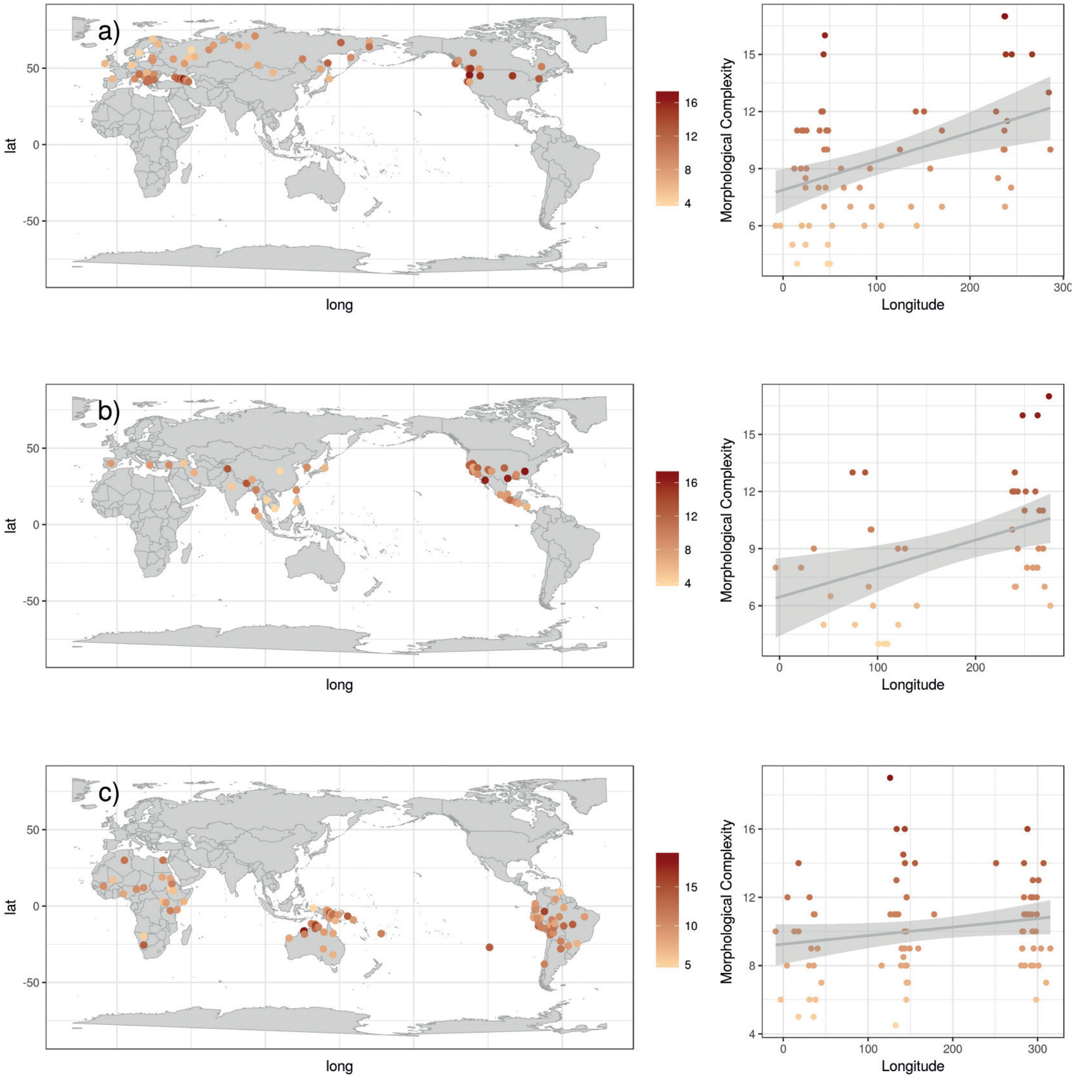
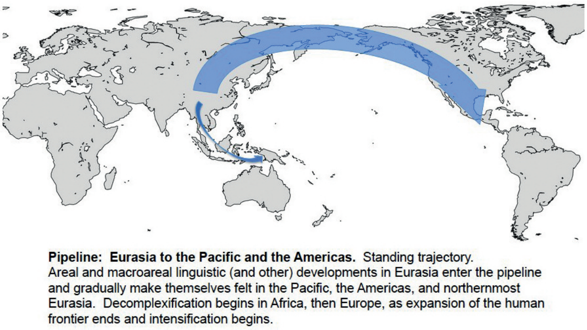


Fig. 5.
Morphological complexity (inventory complexity after Autotyp) plotted against longitude, for three latitudinal bands, and overall 193 languages. Longitudes run from -30 to 330. (a), 65 languages of the northern hemisphere above 40°N ($r = 0.42$, $p < 0.001$); (b), 45 languages of the northern hemisphere below 40° N ($r = 0.42$, $p < 0.01$); (c) 82 languages of the southern continents (Africa at left, Australia-New Guinea-Oceania in center, South America at right) ($r = 0.19$, $p = 0.25$). We here use the Bonferroni method to control for multiple testing, i.e., the original p-value of the Pearson correlation is multiplied by the number of tests (3 in this case). Gray lines indicate linear regression models with 95% confidence bands.

causative alternation, and noun-based vs. verb-based word formation (Nichols, unpublished data). These are grammatical variables not directly related to our complexity measures but sharing a similar broad type of geographical distribution. For all of them, the linguistic population of the Americas generally belongs typologically with the easternmost end of the gradient (cf. Bickel and Nichols 2006). The interpretation of this distribution would appear to be that high morphological complexity has long been (as it is now) a trait of the North Pacific Rim population, from which ancestral colonizing languages entered North America, eventually to populate the entire hemisphere (see Fig. 6). Subsequently, probably beginning only with the Neolithic, complexity levels were reduced in much of the Old World as complex societies and states expanded and trade languages formed, spreading languages to L2 learners. Highland languages, such as the ones surveyed here in the Andean region of South America, might have a) complexified in sociolinguistic isolation, or b) reverted to the earlier and probably default situation (Bickel and Nichols 2003)—or both. Furthermore, in the Americas, the same processes as observed in Eurasia occur independently (at least pre-contact): statehood lowers complexity levels (in Central America and the Andes) compared to more isolated, agricultural highland languages.



Pipeline: Eurasia to the Pacific and the Americas. Standing trajectory. Areal and macroareal linguistic (and other) developments in Eurasia enter the pipeline and gradually make themselves felt in the Pacific, the Americas, and northernmost Eurasia. Decomplexification begins in Africa, then Europe, as expansion of the human frontier ends and intensification begins.

Fig. 6.
Long-standing trajectories of language spread from Eurasia to the Americas (and the Pacific; see also Rootsi et al. 2007; Bickel and Nichols 2006; Nichols 2000).

CONCLUSIONS

Our preliminary analyses of relationships between altitude, longitude and morphological complexity confirm the uncontroversial origin of the indigenous American languages in the North Pacific Rim population. More intriguingly, they indicate that—across the board—high morphological complexity levels (and other traits) of that population have been a) maintained in offspring populations for many millennia, and b) potentially even enhanced through further isolation in certain areas such as the Andes. Thus, the American linguistic population is old enough that contrasts of complexity between high altitude and low altitude areas have developed, and they are observable now. The rate at which these processes evolve is currently not known. The least we can say is that since it has come to affect entire areal populations of languages (and not just the occasional individual language) it is unlikely to have happened quickly, i.e., within a few hundred years. The modern complexity levels, then, are a window into deep population movement, contact, and isolation. Establishing rates of change for simplification and complexification will help to support hypotheses about the possible earliest entries into the Americas.

Appendix 1.

Opacity was surveyed as a whole-language property summed up from a sample of parts of speech and morphological categories:

Nouns:
- Core grammatical cases
- Gender
- Classifiers
- Possessive forms for 1-2 sg. and pl.

Free pronouns:
- 1-2 persons, singular and plural
- Cases
- Gender

Verbs:
- Subject and object person indexes (1-2-3 persons singular and plural)
- Number marking (separate from person-number indexation)
- TAM paradigms for basic synthetic present and aoristic past (or nearest equivalents)

For those contexts the following were surveyed (presence vs. absence):

- Inflectional classes of category markers (affixes, etc.) (formative flexivity, in the terms of Bickel and Nichols 2007:184–5).
- Membership in inflectional classes motivated vs. arbitrary
- Inflectional classes of stem alternations or changes (stem flexivity, in the terms of Bickel and Nichols 2007:184–5).
- Membership in stem alternation classes motivated vs. arbitrary
- Inherent categories (chiefly, gender in nouns)
- Membership in inherent categories motivated vs. arbitrary
- Coexponence (of survey categories with each other)
- Syncretisms within paradigms
- Syncretisms across paradigms (same category)
- Allomorphy
- Discrepancies of position (within a paradigm for some category)
- Discrepancies of wordhood
- Partial marking
- Multiple marking

The survey was conducted by examining paradigms, descriptions of inflection, and descriptions of word classes in grammars. A full description of the measures, the method, the language sample, and the results is in preparation.

REFERENCES

Adelaar, W. F. H., and P. C. Muysken. 2004. *The Languages of the Andes*. Cambridge: Cambridge University Press.

Bentz, C., D. Alikaniotis, M. Cysouw, and R. Ferrer-i-Cancho. 2017. The entropy of words – learnability and expressivity across more than 1000 languages. *Entropy* 19(6), 275; doi:10.3390/e19060275.

Bentz, C. 2016. The Low-Complexity-Belt: Evidence for large-scale language contact in human prehistory? In *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11*), ed. by S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher, and T. Verhoef. doi:10.17617/2.2248195.

Bentz, C., and D. Alikaniotis. 2016. *The word entropy of natural languages*. arXiv preprint, arXiv:1606.06996.

Bentz, C., and A. Berdicevskis. 2016. Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics (COLING 2016), Osaka, Japan*.

Bentz, C., T. Ruzsics, A. Koplenig, and T. Samardžić. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics (COLING 2016), Osaka, Japan*.

Bentz, C., A. Verkerk, D. Kiela, F. Hill, and P. Buttery. 2015. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE* 10 (6), e0128254. doi: 10.1371/journal.pone.0128254.

Bentz, C., and B. Winter. 2013. Languages with more second language speakers tend to lose nominal case. *Language Dynamics and Change* 3: 1–27.

Bickel, B., J. Nichols, T. Zakharko, A. Witzlack-Makarevich, K. Hildebrandt, M. Riessler, L. Bierkandt, F. Zúñiga, and J. B. Lowe. 2017. *The Autotyp database*, release version 0.1.0. https://github.com/autotyp/aut

Bickel, B., and J. Nichols. 2007. Inflectional morphology. In *Language Typology and Syntactic Description*, 3, ed. by Tim Shopen, pp. 169–240. Cambridge: Cambridge University Press.

Bickel, B, and J. Nichols. 2006. Oceania, the Pacific Rim, and the theory of linguistic areas. In *Proceedings of the 32nd Annual Meeting: Special Session on the Languages and Linguistics of Oceania*, ed. by Z. Antić, C. B. Chang, C. S. Sandy, and M. Toosarvandani, pp. 3–15. Berkeley: Berkeley Linguistics Society.

Bickel, B., and J. Nichols. 2003. Typological enclaves. 5th biannual conference, Association for Linguistic Typology.

Bickel, B., and J. Nichols. 2002. T*he Autotyp research program*. Online at: www.autotyp.uzh.ch.

Caron, B. 2013. Hausa grammatical sketch. In *The CorpAfroAs Corpus: A corpus for Afroasiatic languages*, ed. by A. Mettouchi, M. Vanhove, and D. Caubet. Online at http://corpafroas.tge-adonis.fr/Archives/HAU/PDF/. Accessed Feb. 28, 2017.

Corbett, G. 1991. *Gender*. Cambridge: Cambridge University Press.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: Benjamins.

Dobrushina, N. 1999. Glagol. In *Èlementy caxurskogo jazyka v tipologicheskom osveschenii*, ed. by A. E. Kibrik. Moscow: MGU.

Donohue, M., and J. Nichols. 2011. Does phoneme inventory size correlate with population size? *Linguistic Typology* 15.2: 161–170.

Dryer, M. S., and M. Haspelmath, eds. 2013. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. http://wals.info/.

Harris, A. C. 2002. *Endoclisis and the origins of Udi morphosyntax*. Oxford: Oxford University Press.

Jäger, G. 2015. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences* 112(41): 12752–12757.

Karlsson, F., M. Miestamo, and K. Sinnemäki. 2008. Introduction: The problem of language complexity. In *Language Complexity: Typology, Contact, Change*, ed. by M. Miestamo, K. Sinnemäki, and F. Karlsson, pp. vii–xiv. Amsterdam: Benjamins.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *MT summit*, Vol. 5, pp. 79–86.

Koplenig, A., P. Meyer, S. Wolfer, and C. Mueller-Spitzer. 2017. The statistical tradeoff between word order and word structure: Large-scale evidence for the principle of least effort. *PloS ONE* 12.3: e0173614.

Lupyan, G., and R. Dale. 2010. Language structure is partly determined by social structure. *PloS ONE* 5 (1): e8559.

Mayer, T., and M. Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, ed. by N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, pp. 3158–3163. European Language Resources Association (ELRA), Reykjavik, Iceland, May 26-31, 2014.

McWhorter, J. 2007. *Language interrupted: Signs of non-native acquisition in standard language grammars*. Oxford University Press.

McWhorter, J. 2016. Is radical analyticity normal?. In *Cyclical Change Continued*, ed. by E. van Gelderen. Philadelphia/ Amsterdam: John Benjamins Publishing Company.

Newman, P. 2000. *The Hausa language: An encyclopedic reference grammar*. New Haven: Yale University Press.

Nichols, J. 2016. Complex edges, transparent frontiers: Grammatical complexity and language spreads. In *Complexity, Isolation, and Variation*, ed. R. Baechler, and G. Seiler, pp. 117–138. Berlin: de Gruyter.

Nichols, J. 2015. Complexity as non-canonicality. SLE Workshop on Morphological Complexity, Leiden, Sept. 14.

Nichols, J. 2013. The vertical archipelago: Adding the third dimension to linguistic geography. In *Space in Language and Linguistics*, ed. by P. Auer, M. Hilpert, A. Stukenbrock, and B. Szmrecsanyi, pp. 38–60. Berlin: Mouton de Gruyter.

Nichols, J. 2009. Linguistic complexity: A comprehensive definition and survey. In *Language Complexity as an Evolving Variable*, ed. by G. Sampson, D. Gil, and P. Trudgill, pp. 110–125. Oxford: Oxford University Press.

Sinnemäki, K. 2011. Language universals and linguistic complexity: Three case studies in core argument marking. Ph.D. dissertation, University of Helsinki.

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

Van Gijn, R., P. Ranacher, and P. Muysken. n.d. River thinking: River-based diffusion of linguistic features in the Amazon region. (forthcoming)