Language structure is influenced by the proportion of non-native speakers: A reply to Koplenig (2019)

Henri Kauhanen, Sarah Einhaus & George Walkden

Department of Linguistics, University of Konstanz, Konstanz, Germany

May 20, 2022

Abstract

A recent quantitative study claims language structure to be unaffected by the proportion of those speaking the language non-natively [A. Koplenig, *Royal Society Open Science*, 6, 181274 (2019)]. This result hinges on the imputation of an assumed zero proportion of L2 (second-language) speakers for languages which are considered non-vehicular but for which no direct estimate of that proportion exists. We provide an alternative analysis and interpretation of the same data, treating uncertain non-vehicular languages as missing data points. In a multiple regression, controlling for population size, we find that higher proportions of L2 speakers predict lower morphological as well as information-theoretic complexity.

Keywords: language complexity, second-language learning, linguistic typology

1 Introduction

Recent years have seen increased interest in the question of how linguistic structure is influenced by social, historical, environmental and demographic factors. Trudgill [1] makes the case that whether a language complexifies or simplifies over time is affected by who is acquiring it and when: contact situations characterized by a high degree of short-term adult second-language (L2) acquisition are likely to lead to simplification, while contact situations characterized by long-term child bilingualism are likely to lead to additive complexification. Similarly, according to the Linguistic Niche Hypothesis [2], languages with large numbers of speakers, used for communication between strangers, are likely to undergo structural simplification, while languages with smaller numbers of speakers, rarely used for communication between strangers, are likely to develop more structural complexity. In both [1] and [2], adult L2 acquisition is considered to be the key causal factor. This idea has since given rise to substantial amounts of typological [3, 4], corpus-based [5, 6], and experimental [7, 8] research.

In a recent large-scale typological study, on the other hand, Koplenig [9] reports that language complexity—whether quantified in morphological or information-theoretic terms—is not influenced by the share of L2 learners in the speaker population. This conclusion is arrived at through multiple statistical analyses, using either the estimated proportion of L2 speakers,[1] or a categorical variable called "vehicularity", as the predictor. A main aim is to control for the overall absolute number of speakers of a language, considered a potential confound. As Koplenig rightly notes, prior research on the relationship between complexity and proportion of L2 speakers has either used absolute number of speakers as a proxy (e.g. [2]) or has focused on specific linguistic features across a small sample of languages (e.g. [3]), so a direct typological test of this relationship is a desideratum, and this is what Koplenig aims to deliver. He finds that neither the proportion of L2 speakers nor vehicularity predict structure, but that the logarithm of the number of speakers (population size) does.

In this brief contribution, we call some of these conclusions into question. Our main point of contention has to do with the methodological decision in [9] to assign a zero proportion of L2 speakers to non-vehicular languages whenever a direct estimate of that proportion is unavailable. We regard this as an unwarranted researcher degree of freedom, and advocate an alternative analysis: whenever a direct estimate of the proportion of L2 speakers is not available, the language ought to be treated as a missing data point. With this assumption, and using Koplenig's original dataset, we then show that both morphological complexity and information-theoretic complexity are lower for languages with higher proportions of L2 speakers, when population size is controlled for in a multiple regression.

In all that follows, we adopt Koplenig's definitions of morphological and information-theoretic complexity without modification. Morphological complexity is defined over a set of relevant World Atlas of Language Structures (WALS) features [10] and normalized to range between 0 and 1. For example, languages with higher numbers of genders score higher complexities than languages with smaller gender systems (or no gender at all). Information-theoretic complexity is operationalized using the notion of entropy rate [11] and has units of bits per character. In essence, this variable measures redundancy by way of signal compressibility: more redundant constructions lead to better compressibility and hence to lower information-theoretic

[1] The proportion of L2 speakers of a language is simply the number of L2 speakers divided by the sum of the numbers of L1 and L2 speakers. It thus ranges between the extremes of 0 and 1.

complexity. For a more detailed explanation of the two kinds of complexity measures, we refer the reader to Section 2 of [9].

2 Problems with Koplenig's analysis

Estimates of how many L2 speakers a language has are notoriously difficult to come by. Although Koplenig's [9] data source, the 20th edition of Ethnologue [12], provides such numerical estimates for some languages, they are in the minority. To alleviate this problem of data sparsity, the approach adopted in [9] is to make indirect inferences about L2 speaker proportion by way of a notion of linguistic *vehicularity*. Based on Ethnologue's Expanded Graded Intergenerational Disruption Scale (EGIDS; see [13]), Koplenig's vehicularity is a categorical (binary) variable that indicates whether a language 'is used as an L2 in addition to being used as an L1' [9, p. 3], regardless of whether an actual estimate of the number of L2 speakers is available. All non-vehicular languages in the sample for which no such numerical estimate exists are then imputed a zero proportion of L2 speakers. In a series of statistical tests, Koplenig finds that neither vehicularity nor (imputed) proportion of L2 speakers in general predicts complexity, whether understood in the morphological or the information-theoretic sense, challenging previous empirical findings to the contrary based on smaller samples of languages [3].

Although the desire to move in the direction of larger samples strikes us as sensible, we argue that the specific strategy of using vehicularity to impute a large number of zero L2 proportion languages incurs a number of problems, questioning whether the resulting sample is actually representative of reality (on this point, see also [4]). We consequently use neither vehicularity nor imputed L2 speaker proportions in our own analysis. Our reasoning will be explained next, followed by the alternative analysis in Section 3.

First and foremost, a considerable number of non-vehicular languages are reported by Ethnologue to be used as an L2 even though no numerical estimate of L2 users is given and hence such languages appear with a zero proportion of L2 speakers in Koplenig's dataset. For instance, Bawm[2] is noted to be used as an L2 by speakers of Pangkhua, Jaqaru by speakers of Chincha Quechua, Mikasuki by speakers of Muskogee, Low Saxon by speakers of Northern Frisian, and Southern Uzbek by speakers of Turkmen, although no numerical estimates of L2 speaker proportion are given. To assess how often this occurs, we cross-checked Koplenig's sample

[2] All references to Ethnologue in this paper concern its 20th edition [12], the version employed in Koplenig's original study. We employ language names given in Ethnologue; correspondence between Ethnologue and Koplenig's sample was established through the languages' ISO 639-3 codes.

against Ethnologue. The sample contains 1,824 non-vehicular languages with a zero proportion of L2 speakers.[3] In only four cases does Ethnologue provide a numerical zero proportion estimate; thus for 1,820 languages the zero proportion of L2 speakers in [9] has been imputed through non-vehicularity. Of these, Ethnologue however reports that the language is used as an L2 by speakers of some other language or languages in 404 cases (even though no numerical estimate is given). In other words, Koplenig's data imputation is wrong in at least $404/1820 \approx 22\%$ of cases based on this observation alone.

Secondly, a number of non-vehicular languages included in Koplenig's data analysis are considered by Ethnologue to be extinct; they have no known L1 speakers. Examples include (but are not limited to) Atsugewi, Mogholi, Madugele, Yawuru and Quileute. Yet, since Koplenig obtains his population size estimates from a different source [14], these languages feature in his analyses (for instance, Atsugewi with a population of three L1 and zero L2 speakers). We do not think it is sensible to include extinct or near-extinct languages in the same analysis with languages whose speaker numbers are orders of magnitude greater. In some previous research, the methodological decision of imputing a fixed number of 50 speakers for languages whose population size estimates are lower than 50 has been adopted [2, 4]; by contrast, we advocate treating such languages either as missing data or treating them in a separate analysis.

Thirdly, while artificial L2 speaker proportions are imputed in [9] for non-vehicular languages, similar data imputation is not carried out in the case of vehicular languages with missing L2 speaker proportion estimates; for the latter, L2 speaker proportions appear as missing values in Koplenig's sample. This seems unmotivated—if values are imputed for one type of language, why are they not imputed for the other? Even if there was a valid reason for carrying out data imputation for non-vehicular languages only, there are arguably better ways of doing so than to assume all of these languages to have exactly zero L2 speakers. This is because those non-vehicular languages for which estimates *are* available attest L2 speaker proportions which are in many cases considerably different from zero (Figure 1). A less biased imputation scheme would take this structure of the known subset of the entire dataset into account in some principled way [15].

Although we believe the above three reasons cast sufficient doubt over the composition of Koplenig's sample, there are further issues. For instance, the sample has an imputed zero L2 speaker proportion for Karok and Eastern Balochi, even though

[3] The entire sample consists of 2,143 languages, although not every variable has a value for every language, as we discuss in more detail below. In any case, non-vehicular languages with a zero L2 proportion represent the vast majority of the original sample.

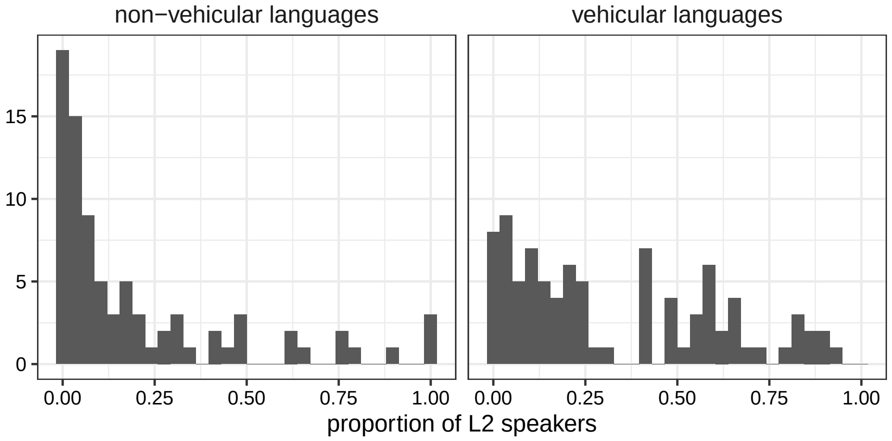

Figure 1. Distribution of the proportion of L2 speakers after uncertain non-vehicular languages have been removed from the dataset.

Ethnologue reports 30 speakers of "some L2 fluency" for the former and notes that the latter has "L2 users of all Balochi languages". Khwedam, also with an imputed zero L2 speaker proportion in Koplenig's sample, is reported by Ethnologue to be learnt by many non-Khwedam speakers for purposes of interaction with speakers of Khwedam. Koplenig's data imputation also fails to take notice of such demographic complexities as are evinced by speakers of Yanyuwa, who according to Ethnologue intermarry with speakers of Garrwa or Mara, with children growing up speaking their mother's language but boys at puberty learning also their father's language. Furthermore, a number of non-vehicular languages with an imputed zero L2 proportion in Koplenig's sample are considered by Ethnologue to have some number of semi-speakers, that is to say, speakers of a dying language who only have limited fluency in the language [16]. Examples include Kutenai, Kwakiutl, Lillooet, Manchu and Mono, though there are many more; an extreme case is Achumawi, which is considered by Ethnologue to have *only* semi-speakers or passive speakers. Although the role of semi-speakers in processes of language simplification or complexification is unclear and deserves further study, we regard the inclusion of languages with significant numbers of semi-speakers (in relation to their numbers of L1 speakers) in the sample as problematic.

Since little reason thus exists to believe that the imputation procedure results in unbiased data, we submit that the safer alternative is to simply regard any language for which the L2 speaker proportion is unknown as a missing data point. When uncertain non-vehicular languages are removed, we are left with a far smaller dataset.[4] As

[4] 171 languages, to be exact. However, the complexity measures are not available for all languages. Morphological complexity is available for 148 languages and information-theoretic complexity for 94 languages.

we show next, however, even this smaller (but more representative) sample is large enough to warrant meaningful statistical inferences.

3 Alternative analysis

Consider a linear model with morphological complexity as the dependent variable and proportion of L2 speakers and log-transformed population size as predictors. When fitted to Koplenig's dataset after removal of uncertain non-vehicular languages, the results are as in Table 1, illustrated graphically in Figure 2A–B. Both population size and the proportion of L2 speakers have a declining effect on morphological complexity, and both predictors are statistically significant.[5] Table 2 reproduces this same analysis, but restricts the dataset to those languages for which at least six WALS features are available for the determination of morphological complexity; here we follow Koplenig who conducts a similar subanalysis to mitigate the problem of the morphological complexity measure being noisy when only very few features are available. The same pattern emerges: higher L2 speaker proportions and larger population sizes both predict lower morphological complexity.[6]

Turning now to information-theoretic complexity, we present results of running the same model, except that this time information-theoretic complexity is the dependent variable. The regression coefficients are listed in Table 3 and illustrated in Figure 2C–D. The proportion of L2 speakers again has a negative effect, but population size has a positive effect. In other words, larger populations appear to attest higher entropy rates, i.e. higher information-theoretic complexity.[7]

4 Discussion

These results suggest that second-language (L2) learning does play a role in linguistic simplification dynamics: higher proportions of non-native speakers imply lower morphological and information-theoretic complexity. On the other hand, population size appears to affect the two measures in different ways: increasing population size decreases morphological but increases information-theoretic complexity. This is an intriguing finding from the point of view of the expectation [9] that morphological and

[5] We do not include an interaction term between the two covariates, as doing so leads to a worse model. The AIC for the interactionless model is 25.66, while the AIC for the model with the interaction is 25.91.

[6] For this analysis, too, addition of an interaction term between the covariates leads to a worse model in terms of AIC: −22.80 (no interaction) vs. −20.85 (with interaction).

[7] There is again no evidence for an interaction between the two covariates, in terms of AIC: 15.95 (no interaction) vs. 17.52 (with interaction).

Table 1. Coefficient estimates for a linear model with morphological complexity as dependent variable and the proportion of L2 speakers and log-transformed population size as independent variables, applied to the dataset without uncertain non-vehicular languages ($N = 148$; $R^2 = 0.11$).

	Estimate	S.E.	t	Pr(> $\lvert t \rvert$)
(intercept)	0.845	0.071	11.92	$< 10^{-15}$ ***
prop. of L2 speakers	−0.278	0.080	−3.48	0.0007 ***
log(population)	−0.014	0.005	−2.99	0.0034 **

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 2. Coefficient estimates for a linear model with morphological complexity as dependent variable and the proportion of L2 speakers and log-transformed population size as independent variables. Restricted to those languages for which the estimation of morphological complexity is possible over at least six features ($N = 101$; $R^2 = 0.14$).

	Estimate	S.E.	t	Pr(> $\lvert t \rvert$)
(intercept)	0.778	0.069	11.28	$< 10^{-15}$ ***
prop. of L2 speakers	−0.251	0.078	−3.21	0.0018 **
log(population)	−0.013	0.005	−2.95	0.0041 **

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 3. Coefficient estimates for a linear model with information-theoretic complexity as dependent variable and the proportion of L2 speakers and log-transformed population size as independent variables, applied to the dataset without uncertain non-vehicular languages ($N = 94$; $R^2 = 0.16$).

	Estimate	S.E.	t	Pr(> $\lvert t \rvert$)
(intercept)	1.344	0.161	8.37	$< 10^{-12}$ ***
prop. of L2 speakers	−0.242	0.10	−2.45	0.0163 *
log(population)	0.027	0.010	2.67	0.0091 **

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

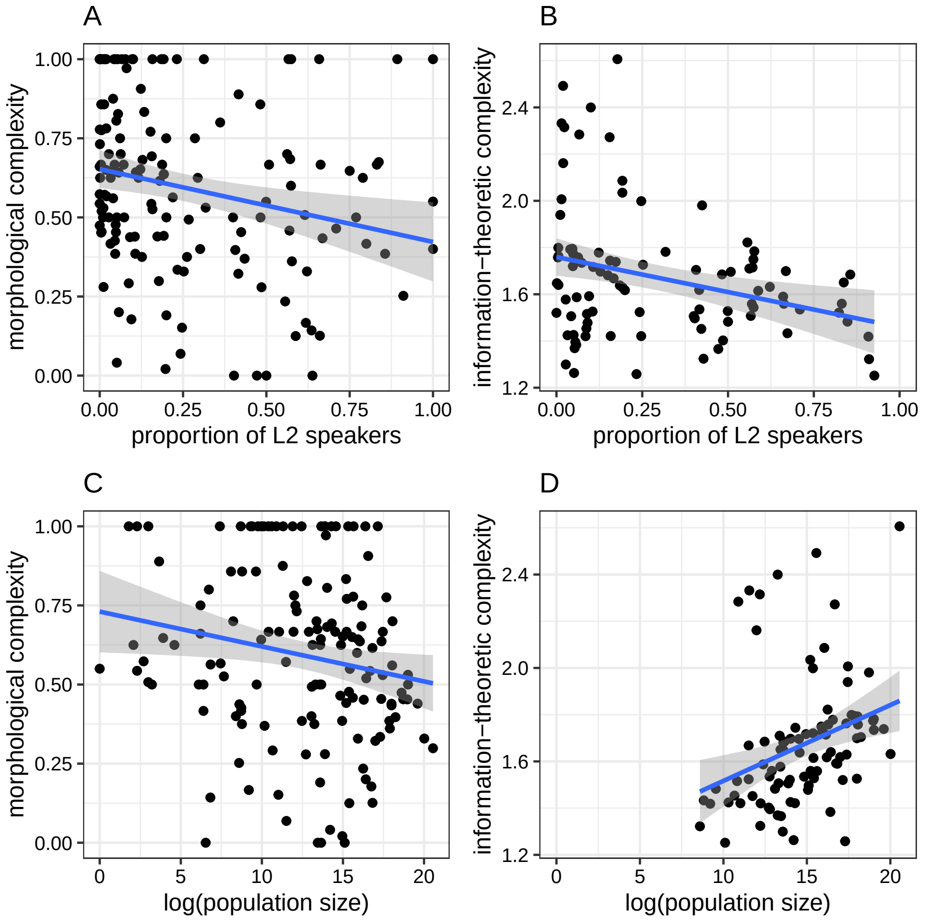

Figure 2. Visualization of the regressions. Morphological (A) and information-theoretic (B) complexity decreases with increasing proportion of L2 speakers. Morphological complexity decreases with increasing population size (C); information-theoretic complexity increases with increasing population size (D).

information-theoretic complexity ought to be inversely correlated: decreasing morphological complexity is expected to decrease redundancy, which increases entropy rate, i.e. information-theoretic complexity. It is possible that the effect of population size simply reflects this general relationship, while the effect of L2 learning adds an additional layer on top of it—this is exactly the prediction if L2 learners specifically struggle with both morphological complexity and lack of signal redundancy. It is less clear, at this moment, why larger speaker populations should prefer either lower morphological complexity or lower redundancy; in this respect, there is much room for further research into the mechanisms whereby external factors influence the development of language structure.

To obtain these results, we have treated uncertain non-vehicular languages—in other words, non-vehicular languages for which no direct estimate of the proportion of L2 speakers is available—as missing data. We believe this is no less justified than Koplenig's procedure of filling in the blanks. On a more fundamental level, however, one may wish to ask what it means for a language to have a zero proportion of L2 speakers. We would find it rather surprising if very many natural languages attested no L2 speakers at all. The decision to treat all uncertain non-vehicular languages as zero-L2 languages, and to fold these languages in with the rest for purposes of statistical analysis, therefore strikes us as an unmotivated approximation. Since it places the majority of the original sample singularly at zero proportion, it is unsurprising that a statistical analysis conducted on that full sample struggles to detect an effect of the proportion of L2 speakers on the complexity measures. Noise from the zero-proportion majority clouds the signal from the non-zero-proportion minority.

Our results are in line with those of Sinnemäki [4], who shows for a sample of 66 languages that proportion of L2 speakers is an important predictor of both the number of morphological cases (more L2 speakers correlates with fewer cases) and the probability of having morphological case at all (the more L2 speakers, the lower the probability). Sinnemäki observes that his results are at variance with those of Koplenig [9], and suggests that this discrepancy may be due either to the smaller sample size in his study or to Koplenig's assumption that languages with EGIDS 4 or lower (circa 91% of languages in Ethnologue) have no non-native speakers. The fact that our somewhat larger sample (148 languages), with a different measure of morphological complexity as the dependent variable, yields a similar result to those of [4] suggests that the latter assumption is probably the driving force in the discrepancy.

Even though proportion of L2 speakers and population size have a detectable (statistically significant) effect on two types of language complexity in our findings, we hasten to add that the explained variances remain small: on the order of 10–15% (see the value of R^2 in the captions to Tables 1–3, and the large spread of the data points in Figure 2). This is unsurprising, given the enormous complexity of the socio-historical processes that are responsible for the linguistic variation observable today [17]. The overall predictive power of statistical models of this kind may be increased by considering additional factors such as language family-specific or area-specific random effects. Since our focus here has been solely on the contribution of L2 speaker proportion and population size, we have not considered such additions.

Theoretically, one may wonder to what extent the relationships between variables such as morphological or information-theoretic complexity, vehicularity, proportion of L2 speakers and (the logarithm of) population size are linear at all. For instance, perhaps there does exist a valid distinction between vehicular and non-vehicular languages in the sense that two languages from these two classes may behave qualitatively differently even if their actual proportions of L2 speakers are similar. Similarly, it is unclear to us whether highly vehicular languages with extremely high numbers of speakers, such as English and Spanish, should be expected to behave in general like languages which are more modal in terms of the global distribution of population size. We leave these questions open for future research.

Funding

The work reported here was funded by the European Research Council as part of project STARFISH (851423). Access to Ethnologue was funded by the Communication, Information, Media Centre (KIM) of the University of Konstanz.

Data availability

Data and code (R, version 4.0.4; [18]) may be obtained from https://github.com/erc-starfish/koplenig-reply.

References

[1] Peter Trudgill. *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford: Oxford University Press, 2011.

[2] Gary Lupyan and Rick Dale. “Language structure is partly determined by social structure”. In: *PLoS ONE* 5 (2010), e8559. doi: 10.1371/journal.pone.0008559.

[3] Christian Bentz and Bodo Winter. “Languages with more second language learners tend to lose nominal case”. In: *Language Dynamics and Change* 3 (2013), pp. 1–27. DOI: 10.1163/22105832-13030105.

[4] Kaius Sinnemäki. “Linguistic system and sociolinguistic environment as competing factors in linguistic variation: a typological approach”. In: *Journal of Historical Sociolinguistics* 6 (2020), p. 20191010. DOI: 10.1515/jhsl-2019-1010.

[5] Katharina Ehret and Benedikt Szmrecsanyi. “Compressing learner language: an information-theoretic measure of complexity in SLA production data”. In: *Second Language Research* 35 (2019), pp. 23–45. DOI: 10.1177/0267658316669559.

[6] George Walkden and Anne Breitbarth. “Complexity as L2-difficulty: implications for syntactic change”. In: *Theoretical Linguistics* 45 (2019), pp. 183–209. DOI: 10.1515/tl-2019-0012.

[7] Mark Atkinson, Kenny Smith, and Simon Kirby. “Adult learning and language simplification”. In: *Cognitive Science* 42 (2018), pp. 2818–2854. DOI: 10.1111/cogs.12686.

[8] Aleksandrs Berdicevskis and Arturs Semenuks. “Imperfect language learning reduces morphological overspecification: experimental evidence”. In: *PLoS ONE* 17 (2022), e0262876. DOI: 10.1371/journal.pone.0262876.

[9] Alexander Koplenig. “Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers”. In: *Royal Society Open Science* 6 (2019), p. 181274. DOI: 10.1098/rsos.181274.

[10] Matthew S. Dryer and Martin Haspelmath, eds. *The World Atlas of Language Structures Online*. Leipzig, 2013. URL: http://wals.info/.

[11] Alexander Koplenig et al. “The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort”. In: *PLoS ONE* 12 (2017), e0173614. DOI: 10.1371/journal.pone.0173614.

[12] Gary F. Simons and Charles D. Fennig, eds. *Ethnologue: Languages of the World*. 20th edition. Dallas, TX: SIL International, 2017. URL: http://www.ethnologue.com.

[13] M. Paul Lewis and Gary F. Simons. “Assessing endangerment: expanding Fishman’s GIDS”. In: *Revue Roumaine de Linguistique* 55 (2010), pp. 103–120.

[14] Tatsuya Amano et al. "Global distribution and drivers of language extinction risk". In: *Proceedings of the Royal Society B* 281 (2014), p. 20141574. DOI: 10.1098/rspb.2014.1574.

[15] Craig K. Enders. *Applied missing data analysis*. New York, NY: The Guilford Press, 2010.

[16] Nancy C. Dorian. "The problem of the semi-speaker in language death". In: *Linguistics* 15 (1977), pp. 23–32. DOI: 10.1515/ling.1977.15.191.23.

[17] Daniel Nettle. *Linguistic diversity*. Oxford: Oxford University Press, 1999.

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.