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

2023

This is a pre-copyedited, author-produced version of an article accepted for publication in *Journal of Language Evolution* following peer review. The version of record is available online at https://doi.org/10.1093/jole/lzad005.

Acknowledgements

The work reported here was funded by the European Research Council as part of project STARFISH (851423). Access to Ethnologue was made possible by the Communication, Information, Media Centre (KIM) of the University of Konstanz. We thank Frederik Hartmann for discussions, as well as the reviewers for their constructive criticisms and fruitful suggestions, and especially for encouraging us to pursue the missing data and imputation problem further than we had originally set out to do.

Data availability

Data and code may be obtained from https://doi.org/10.5281/zenodo.7752933.

*Corresponding author: henri.kauhanen@uni-konstanz.de. Department of Linguistics, P.O. Box 226, University of Konstanz, 78457 Konstanz, Germany.

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

2023

Abstract

A recent quantitative study claims language structure, whether quantified as morphological or information-theoretic complexity, 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 either the use of a categorical notion of 'vehicularity' as a proxy for the proportion of L2 (second-language) speakers, or the imputation of an assumed zero proportion of L2 speakers for languages which are considered non-vehicular but for which no direct estimate of that proportion exists. We provide two alternative analyses of the same data. The first reanalysis treats uncertain non-vehicular languages as missing data points; the second one employs multiple imputation to fill in the missing data. Mixed effects models find a statistically significant negative relationship between proportion of L2 speakers and morphological complexity: in both reanalyses, a higher proportion of L2 speakers predicts lower morphological complexity. We find no statistically significant evidence for a relationship between proportion of L2 speakers and information-theoretic complexity, however.

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

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 (2011) 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 (Lupyan and Dale, 2010), 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 Trudgill (2011) and Lupyan and Dale (2010), adult L2 acquisition is considered to be the key causal factor in simplification. This idea has since given rise to substantial amounts of typological (Bentz and Winter, 2013; Sinnemäki, 2020), corpus-based (Ehret and Szmrecsanyi, 2019; Walkden and Breitbarth, 2019), and experimental (Atkinson et al., 2018; Berdicevskis and Semenuks, 2022) research.

In a recent large-scale typological study, on the other hand, Koplenig (2019) 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 or a categorical variable called "vehicularity" as a 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. Lupyan and Dale, 2010) or has focused on specific linguistic features across a small sample of languages (e.g. Bentz and Winter, 2013), 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 paper, we call some of these conclusions into question. Our main point of contention has to do with the methodological decision in Koplenig (2019) to use vehicularity as a proxy for the number of L2 speakers and the subsequent step of imputing a zero proportion of L2 speakers to non-vehicular languages whenever a direct estimate of that proportion is unavailable. We explore the merits of two alternative reanalyses of Koplenig's original dataset: a conservative one in which missing values are not imputed at all, and another in which imputation is carried out more systematically using the technique of multilevel multiple imputation (van Buuren, 2018). Both analyses point to the same conclusion with regard to the relationship between proportion of L2 speakers and morphological complexity: unlike in Koplenig's original analysis, morphological complexity is lower for languages with higher proportions of L2 speakers. For information-theoretic complexity, on the other hand, we find no such dependency on the proportion of L2 speakers, so that here our results align with those in Koplenig (2019).

The paper is structured as follows. Section 2 discusses linguistic complexity, specifically, the two complexity measures studied in Koplenig (2019); Section 3 presents our critique of vehicularity and zero-imputation as implemented in that study. In Section 4, we discuss the problem of missing data and imputation on a conceptual level, suggesting that while zero-imputation is problematic, other types of imputation may in fact be fruitfully used to deal with missing data. Section 5 presents our alternative analyses. Section 6 summarizes our findings, discusses the difference observed between morphological and information-theoretic complexity in more detail, and expands on the challenges involved in dealing with missing data in linguistic typology.

2 Linguistic complexity

The notion of complexity in language can be understood in many different ways: see e.g. Dahl (2004, ch. 4), Miestamo (2008), and Walkden and Breitbarth (2019) for discussion. Koplenig (2019) takes his lead from Lupyan and Dale (2010) in defining and assessing two types of complexity: morphological complexity and information-theoretic complexity.

Morphological complexity is defined by Koplenig (2019) based on 28 features from the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). Each feature can have more than one value, e.g. 'Number of genders' can be 'None', 'Two', 'Three', 'Four', or 'Five or more'. These values are mapped to an ordinal scale from 1–5, and analogously for other features; this scale is then normalized to the interval $[0, 1]$ for all features, 0 being maximally simple and 1 being maximally complex. The overall morphological complexity of a language is then defined as the mean of the normalized complexity scores for individual features: if n_f features are recorded for a language and c_i is the normalized complexity of the ith feature, then the language's overall

morphological complexity C is calculated as

$$C = \frac{1}{n_f} \sum_{i=1}^{n_f} c_i. \tag{1}$$

Thus C is also constrained to the interval $[0, 1]$, with higher values indicating greater complexity. This measure is discussed and motivated in more detail in Bentz et al. (2016). Since not all languages in WALS have values for all features, Koplenig (2019) runs separate analyses on the full dataset (i.e. languages with values for at least one feature) and a subset of languages with values for at least six features. The latter is approximately half the size of the former.

Koplenig (2019) derives his information-theoretic complexity metric from corpus data, specifically the Gospel of Mark in over a thousand different languages. The intuition is based on estimating the entropy rate of a stationary random source producing texts (sequences of characters): a text with more redundancy is more compressible, and hence has a lower information-theoretic complexity. Compressibility is operationalized based on the non-parametric entropy estimation method of Kontoyiannis (1997), and has units of bits per character; the specific entropy rate estimates are taken from Koplenig et al. (2017), who also provide more detail on the method. To determine the amount of redundancy at a given text position i (the ith character in a sequence of M characters), one assesses how many of the characters starting at i have already occurred in the same order (as a substring) previously in the text; the quantity of interest is the longest match length plus one, notated l_i. Longer match lengths are indicative of more redundancy and hence of lower entropy; formally, it has been shown that as the amount of text observed grows to infinity, the quantity $l_i/\log(i)$ converges to the inverse of entropy, H^{-1} (Ornstein and Weiss, 1993). For a finite text, to make maximal use of the available data, H is estimated by taking a growing-window average over all text positions i:

$$\widehat{H} = \left[\frac{1}{M} \sum_{i=2}^{M} \frac{l_i}{\log_2(i)} \right]^{-1} \tag{2}$$

where the logarithm is taken to base 2 to yield entropy in units of bits.

The rationale for including information-theoretic complexity in the first place comes from Lupyan and Dale (2010), who claim that morphological overspecification (i.e. morphological complexity) correlates with redundancy, and that redundancy facilitates child (but not adult) language acquisition by providing multiple cues to the same information; redundancy is therefore, in their account, the driving force behind differences between child and adult language acquisition. In their supplementary materials (Lupyan and Dale, 2010, Text S11) they operationalize redundancy by assessing the compressibility (using WinRAR) of translations of the same text in 103 languages, finding a correlation between higher population size and lower compressibility. As Koplenig (2019) observes, lower compressibility can also be viewed as greater complexity from an information-theoretic perspective, since low compressibility means low redundancy and hence it is more difficult to predict following material:

> Thus, the results of [Lupyan and Dale (2010)] indicate that languages with more speakers are morphologically less complex, but informationally more complex. This points towards a negative statistical association between morphological complexity and entropy rates (Koplenig, 2019, 2).

However, Koplenig's own analysis finds no significant monotonic relationship between morphological and information-theoretic complexity once other factors are controlled for (Koplenig, 2019, 7–8), a fact which calls Lupyan & Dale's explanatory hypothesis into question.

For a more in-depth discussion of the two kinds of complexity measures, we refer the reader to Section 2 of Koplenig (2019) as well as the references therein. In all that follows, we adopt Koplenig's definitions of morphological and information-theoretic complexity without modification.

3 Critique of Koplenig's analysis

Estimates of how many L2 speakers a language has are notoriously difficult to come by. Koplenig's data source, the 20th edition of Ethnologue (Simons and Fennig, 2017a), provides such numerical estimates for some languages—but they are in the minority. To alleviate this problem of data sparsity, the approach adopted in Koplenig (2019) 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 Lewis and Simons 2010), Koplenig's vehicularity is a categorical (binary) variable that intends to indicate whether a language 'is used as an L2 in addition to being used as an L1' (Koplenig, 2019, 3), regardless of whether an actual estimate of the number of L2 speakers is available. The EGIDS is a 10-point scale used by Ethnologue to assess language status, ranging from 0 'international' to 10 'extinct' (see Table 1). In Koplenig's analysis, languages with an EGIDS score of 3 or lower are defined to be vehicular, the rest being non-vehicular.

Vehicularity thus provides a crude estimate of whether a language is likely to have significant numbers of L2 speakers. In Koplenig (2019), vehicularity also serves as a stepping stone to a secondary analysis based on the proportion of L2 speakers: whenever a non-vehicular language in the sample is missing an L2 proportion estimate in Ethnologue,[1] a proportion of 0 is imputed for that language. We call this analysis strategy *zero-imputation* in what follows. 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 (Bentz and Winter, 2013).

Although the desire to move in the direction of larger samples strikes us as sensible, we argue that both strategies—using vehicularity as a binary predictor, and using vehicularity as a means of zero-imputation for the proportion of L2 speakers—are problematic. We consequently use neither vehicularity nor zero-imputation in our own analyses. Our reasoning for this choice will be explained next, followed by a discussion of the missing data problem in Section 4 and our alternative analyses in Section 5.

First and foremost, even though vehicularity was intended as a categorical measure that captures whether L2 speakers exist or not, 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. For instance, Bawm 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.[2] To assess how often this occurs, we cross-checked Koplenig's sample against Ethnologue. The sample contains 1,824 non-vehicular languages with a zero proportion of L2 speakers, the total number of languages in the sample being 2,143. In only four cases does Ethnologue provide an actual numerical zero proportion estimate; thus for 1,820 languages the zero proportion of L2 speakers in Koplenig (2019) 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, the EGIDS-based categorical variable of vehicularity fails to capture the information it was intended to capture—whether L2 speakers exist—in at least $404/1820 \approx 22\%$

[1] To be exact, Ethnologue does not record L2 speaker *proportions* but rather absolute numbers of L1 and L2 speakers. The proportion of L2 speakers is calculated as the ratio of the number of L2 speakers to the sum of the numbers of L1 and L2 speakers.

[2] All references to Ethnologue in this paper concern its 20th edition (Simons and Fennig, 2017a), 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. The 'used as L2' information appears in the 'Language Use' data field in Ethnologue, and is thus separate from EGIDS classification.

Table 1. The EGIDS scale of Ethnologue (Simons and Fennig 2017b; also see Lewis and Simons 2010), together with the mapping to vehicularity (Koplenig, 2019).

Vehicular?	EGIDS	Label	Description
Yes	0	International	The language is widely used between nations in trade, knowledge exchange, and international policy.
Yes	1	National	The language is used in education, work, mass media, and government at the national level.
Yes	2	Provincial	The language is used in education, work, mass media, and government within major administrative subdivisions of a nation.
Yes	3	Wider Communication	The language is used in work and mass media without official status to transcend language differences across a region.
No	4	Educational	The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.
No	5	Developing	The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.
No	6a	Vigorous	The language is used for face-to-face communication by all generations and the situation is sustainable.
No	6b	Threatened	The language is used for face-to-face communication within all generations, but it is losing users.
No	7	Shifting	The child-bearing generation can use the language among themselves, but it is not being transmitted to children.
No	8a	Moribund	The only remaining active users of the language are members of the grandparent generation and older.
No	8b	Nearly Extinct	The only remaining users of the language are members of the grandparent generation or older who have little opportunity to use the language.
No	9	Dormant	The language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency.
No	10	Extinct	The language is no longer used and no one retains a sense of ethnic identity associated with the language.

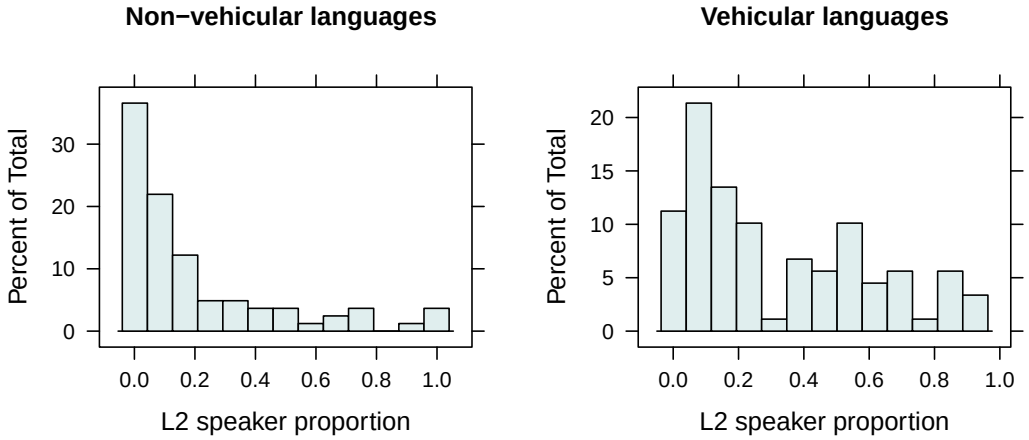

Figure 1. Distribution of the proportion of L2 speakers after uncertain non-vehicular languages have been removed from the dataset.

of cases based on this observation alone.

Quite independently of the above issue, the strategy of zero-imputing L2 proportions based on vehicularity is also problematic. There exists a vast statistical literature on missing data and imputation (see Enders 2010 and van Buuren 2018 for book-length treatments); the imputation of a fixed numerical value (such as zero L2 speaker proportion) is virtually never recommended in this literature, however. Different imputation methods make different assumptions about the underlying process that leads to missing data (see Section 4); it is by no means clear that zero-imputation of L2 speaker proportions, especially when the fraction of missing data is as high as in Koplenig's sample, leads to unbiased regression estimates at the analysis stage. In fact, a glance at the distribution of L2 speaker proportion for those non-vehicular languages for which these estimates are available reveals that these proportions are not limited to ranges near 0; indeed, some non-vehicular languages are reported by Ethnologue to have nothing but L2 speakers (Figure 1).

The decision to base vehicularity on an arbitrary cutoff point on the EGIDS scale also leads to a number of further problems in individual cases. For instance, the sample regards Karok and Eastern Balochi as non-vehicular (and hence as having zero L2 speaker proportions), even though 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. Vehicularity and zero-imputation also fail 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.

In summary, we argue that vehicularity is not a faithful indicator of a language's propensity to be used as an L2, since in as many as 22% of cases Ethnologue in fact reports these languages as having at least some L2 speakers, regardless of the language's EGIDS status and regardless of whether an actual estimate of the number of L2 speakers is given. We have also called into question the suitability of zero-imputation, and will next discuss ways of improving on the missing data problem by utilizing a more systematic imputation method which takes the structure of the known subset of the entire sample into consideration when imputing missing values.

4 Missing data and imputation

Above, we have criticized both the use of vehicularity as a predictor as well as the practice of zero-imputing L2 speaker proportions based on vehicularity. It was pointed out that in 404 cases out of 1,820, Ethnologue reports that a non-vehicular language is used as an L2 by one or more languages even though no numerical L2 speaker proportion estimate is given. It would be tempting to use this information as the basis of a statistical analysis, replacing (or complementing) Koplenig's vehicularity. However, such an analysis is not warranted, for the following reason: the absence of evidence for L2 speakers does not constitute evidence for their absence. This presents us with a difficult missing data problem.

Collecting numerical estimates of proportions of L2 speakers can in fact be seen as a way of tackling that missing data problem: every recorded numerical estimate is by definition a non-missing data point and ready to be analysed as such, as a numerical variable. Although it would perhaps be possible to define some cutoff point in terms of proportion of L2 speakers to separate languages into those which are vehicular and those which are non-vehicular, this strategy introduces an unnecessary researcher degree of freedom to the analysis (much like deciding where to set the cutoff point on the EGIDS scale for the purpose of defining vehicularity). Consequently, we opt to work with the raw proportion estimates instead. Another benefit of this kind of analysis is that no information is lost at the data preparation stage.

We are still faced with a missing data problem, however, since so few languages have a numerical L2 speaker proportion estimate in Ethnologue. There are two possibilities: we can either ignore those languages for which no estimate is available, or we can attempt to fill in the missing estimates in some way. The first approach is known as complete cases analysis, while the latter can be implemented using a number of imputation techniques. The choice between these options should reflect our belief about why the missing data are missing. It is possible for data to be missing completely at random (abbreviated MCAR), which means that the probability of any given data point to go missing is independent of any other aspect of the population (van Buuren, 2018). In this case, complete cases analysis leads provably to unbiased regression estimates, and hence the presence of missing data leads only to a certain loss of statistical power as the sample is smaller than it could, in principle, be. We provide a complete cases analysis of Koplenig's data in Section 5.1.

It is also possible that the probability of a data point going missing is not completely random, but instead depends in some way on a variable other than the variable whose values are being imputed. For instance, it is possible (and in fact quite conceivable) that the probability of an L2 speaker proportion being missing is dependent on population size, with larger languages being better described in typological databases such as Ethnologue. This is known (somewhat confusingly) as a missing at random (MAR) mechanism in the specialist literature (van Buuren, 2018). In this case, simple methods such as complete cases analysis (but also certain popular forms of imputation, such as mean imputation and simple regression imputation, as well as the practice of zero-imputation employed in Koplenig 2019) are not guaranteed to yield unbiased estimates. In other words, in a MAR setting these techniques may skew the analysis and produce results which are not representative of the population. Techniques exist, however, for dealing with MAR data. The state of the art is represented by multiple imputation, in which several imputed copies of the original dataset are created, analysed separately and then pooled together for final regression estimates; since the imputation is carried out multiple times, the uncertainty inherent in the imputation ends up reflected in the variances of the resulting regression estimators at the analysis stage (van Buuren, 2018). We provide a multiple imputation analysis of Koplenig's data in Section 5.2.

Finally, it is possible that the probability of a data point being missing depends on the imputed

variable itself. In this case, we have a missing not at random (MNAR) mechanism. Not much can be done in this case, especially when the proportion of missing data is large. We return to a discussion of the possibility that L2 speaker proportions are systematically missing not at random, and the consequences of this possibility, in Section 6.

5 Alternative analyses

5.1 Reanalysis 1: complete cases analysis

If data are missing completely at random, complete cases analysis is feasible. In this technique, data entries which lack information for a variable (here, the proportion of L2 speakers) are simply left out of the analysis. Doing this with Koplenig's sample, we are left with a total of 171 languages, 148 of which have information about morphological complexity and 94 of which have information about information-theoretic complexity. Although vehicularity plays no role in our analysis, we point out for the sake of completeness that 89 of these 171 languages are vehicular and 82 non-vehicular according to Koplenig's criterion. Our sample for the complete cases analysis includes languages from 29 unique language families and 21 unique linguistic areas. The three most frequent families contain 48.5% of the languages; the three most frequent areas contain 45.8% of the languages.[3]

For purposes of the complete cases analysis, our model of the data is a mixed effects linear regression model with proportion of L2 speakers (a real number between 0 and 1) and the (natural) logarithm of population size as fixed effects. We also include random intercepts for language family and linguistic area whenever this is possible (whenever the models do not incur convergence or singularity issues).[4] Since information about linguistic area is also missing in the dataset for a considerable number of languages (414 to be exact), in some of the analyses the resulting fits are singular when a random intercept is included for linguistic area. In those cases, we only include a random intercept for language family. The dependent variable is either morphological complexity or information-theoretic complexity (see Section 2). We do not include interactions between the covariates in any of our models, as doing so always leads to a worse model when quantified on AIC (see Supplementary Materials). We use R, version 4.0.4 (R Core Team, 2021) and *lme4* (Bates et al., 2015) for the regressions, obtain p-values with *lmerTest* (Kuznetsova et al., 2017) and use *effects* (Fox and Weisberg, 2019) for effects plots.

The results of this complete cases analysis are presented in Table 2 and illustrated graphically in Figure 2A–B for morphological complexity. Both population size and the proportion of L2 speakers have a declining effect on morphological complexity, and both predictors are statistically significant. Table 3 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 (2019) who conducts a similar subanalysis to mitigate the problem of the morphological complexity measure being noisy when only very few features are available. (This analysis does not include a random intercept for linguistic area, as doing so leads to a singular fit.)

[3]These numbers should be contrasted with the corresponding figures for Koplenig's full sample. In the full sample, 126 unique language families and 25 unique linguistic areas are recorded. The three most frequent families have 39.5% of the data, while the three most frequent areas have 32.3%. The complete cases analysis thus discards a great number of the rarer families, when compared to the full dataset. This discrepancy will be greatly reduced in our second analysis, where multiple imputation allows us to retain more families in the dataset (Section 5.2).

[4]It would be ideal to include random slopes as well to test for the possibility that in some families or areas the relation between complexity and the demographic variables goes in the unexpected direction (e.g. increasing L2 speaker proportion predicting higher morphological complexity). However, models where either the slope of L2 speaker proportion or the slope of population size varies by either family or area in general lead to convergence problems. An exception is the regression of morphological complexity, where we can add a random slope for L2 speaker proportion conditioned by language family, but only if no random effect is included for linguistic area. The result aligns with what we report for the random intercepts models below (see Supplementary Materials).

The same pattern emerges: higher L2 speaker proportions and larger population sizes both predict lower morphological complexity.

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. Again, the missingness of the area information for multiple languages precludes the inclusion of a random intercept for linguistic area. The regression coefficients are listed in Table 4 and the effects illustrated in Figure 2C–D. We find no evidence for an effect of either the proportion of L2 speakers or population size on information-theoretic complexity based on this complete cases analysis.

5.2 Reanalysis 2: multiple imputation

The complete cases analysis assumes that L2 speaker proportions are missing completely at random; if this were not the case, then there would be no guarantee that the resulting regression coefficient estimates are unbiased. It is possible for this assumption to be false: the assumption would be falsified if, for instance, languages with more speakers were more likely to have L2 speaker proportions recorded, or if different language families or linguistic areas were recorded for L2 speaker proportion with different propensities. Relaxing these assumptions necessitates the use of imputation techniques which can deal with such 'missing at random' (MAR) data.

In multiple imputation, $m > 1$ copies of the dataset are created and missing values are imputed separately for each copy, yielding *m completed samples*. The practice derives from the intuition that imputing a single value for a given missing data point can never be right—if we had complete certainty in our imputation, the data point would in fact not be missing (van Buuren, 2018). Imputing multiple values, when done appropriately, quantifies the uncertainty inherent in the missing data and leads to unbiased regression estimates under a MAR mechanism of missingness. After the imputation phase, each of the m datasets is subjected to the same regression analysis. After the analysis phase, the regression results from the m completed samples are pooled together using so-called Rubin's rules (see Enders, 2010, 221–224), yielding the final estimates for each predictor.

Several different methods are available for the imputation phase. We use the multilevel multiple imputation tools provided by the *mice* package (van Buuren and Groothuis-Oudshoorn, 2011), using method *l2.lmer* and choosing $m = 100$. A separate imputation model is constructed for the regression of each type of complexity (morphological and information-theoretic); while it would be possible to construct a single imputation model covering both variables in principle, the fact that missing values appear in both response variables and that the attested portions of the data overlap only to a limited extent means that by specifying separate imputation models we can retain more languages for the regressions.[5] Each imputation consists of the application of a linear model with language family as a clustering variable; the predictors we use to impute L2 speaker proportions are complexity (either morphological or information-theoretic), logarithmic population size and logarithmic range size (an estimate of the extent of the geographical area in which a language is spoken); these are the predictors available to us in Koplenig's dataset.[6] Although it would be ideal to include linguistic area among these variables, it turned out that the high missingness rate in

[5]Morphological complexity is available for 1,581 languages and information-theoretic complexity for 1,088 languages, but the number of languages which have both is only 526. We have carried out an additional analysis in which a single imputation model is constructed instead of the two separate models as described below. The results do not change (see Supplementary Materials).

[6]It may seem circular at first sight to include the dependent variables of the ultimate analysis (i.e. morphological complexity and information-theoretic complexity) as predictors when imputing missing values for their ultimate predictor (i.e. L2 speaker proportion). However, this is the recommended course of action in multiple imputation (see Enders, 2010, 201–202).

Table 2. Fixed effect 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, including random intercepts for language family and linguistic area. Complete cases analysis ($N = 144$).

	Estimate	S.E.	t	Pr($>|t|$)
(intercept)	0.817	0.084	9.70	≈ 0 ***
prop. of L2 speakers	−0.243	0.082	−2.95	0.004 **
log(population)	−0.016	0.006	−2.71	0.010 *

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

Table 3. Fixed effect 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, including a random intercept for language family. Complete cases analysis restricted to those languages for which the estimation of morphological complexity is possible over at least six features ($N = 101$).

	Estimate	S.E.	t	Pr($>|t|$)
(intercept)	0.781	0.078	9.95	≈ 0 ***
prop. of L2 speakers	−0.218	0.078	−2.79	0.006 **
log(population)	−0.018	0.006	−3.23	0.002 **

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

Table 4. Fixed effect 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, including a random intercept for language family. Complete cases analysis ($N = 94$).

	Estimate	S.E.	t	Pr($>|t|$)
(intercept)	1.429	0.185	7.72	≈ 0 ***
prop. of L2 speakers	−0.161	0.100	−1.61	0.111
log(population)	0.018	0.012	1.53	0.130

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

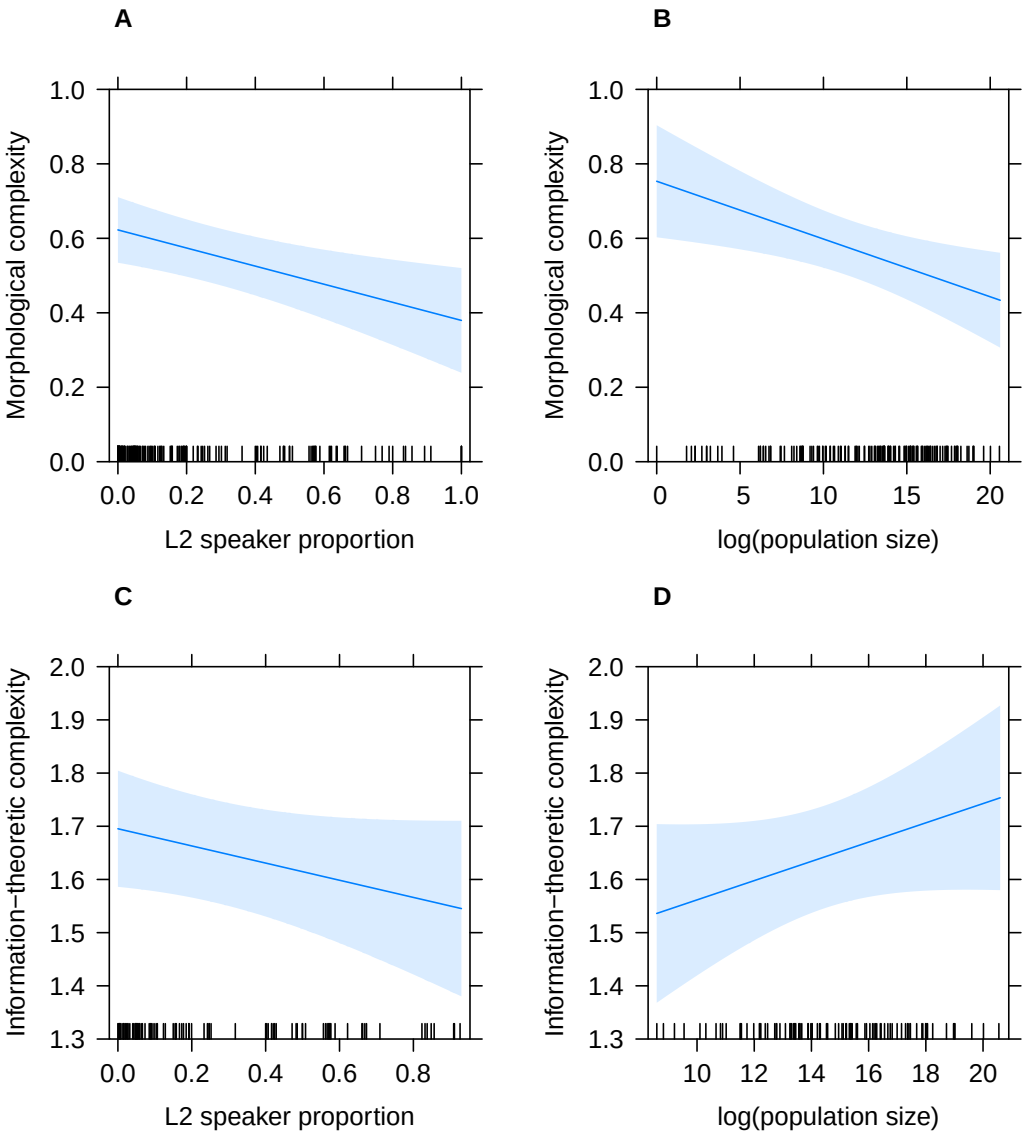

Figure 2. Effects plots (complete cases analysis). Morphological complexity decreases with increasing L2 speaker proportion (A) and with increasing population size (B). There is no statistically significant effect of either L2 speaker proportion or population size on information-theoretic complexity (C–D).

this variable itself precluded the imputation model from running. Although the dataset includes missing values in many variables beyond L2 speaker proportion, we only impute missing values in the L2 speaker proportion variable; our attempts to deal with missingness in many variables simultaneously on multiple levels turned out unsuccessful. The sample sizes of our multiply-imputed regression analyses are nevertheless considerably larger than those of the complete cases analysis: in the case of morphological complexity, we have 1,499 languages in 122 families and 24 areas; in the case of information-theoretic complexity, we have 717 languages spread over 79 families and 23 areas.

An important technical point is that the imputation of physically impossible values should not occur; here, we need to make sure that L2 speaker proportions lower than 0 or higher than 1 are never imputed. To guarantee this, we use a logit transform of the L2 speaker proportion variable to project that variable to the real line:

$$\rho' = \log\left(\frac{\epsilon + (1-2\epsilon)\rho}{1-(\epsilon+(1-2\epsilon)\rho)}\right) \tag{3}$$

where ρ is the original L2 speaker proportion variable, ρ' is the transformed variable, and $\epsilon = 10^{-5}$ is a small positive number used to nudge the limits of the transformation so that data points with $\rho = 0$ or $\rho = 1$ do not disappear into infinity. Thus the imputation model as well as the regression analysis uses ρ' rather than ρ; this is reflected in the units of the resulting coefficients but does not otherwise change the nature of the regressions.

For the regression model, we again include fixed effects for (logit-transformed) L2 speaker proportion and the logarithm of population size, as well as random intercepts for language family and linguistic area. The results, pooled over the $m = 100$ completed copies of the dataset, are summarized in Table 5 for morphological complexity and in Table 6 for information-theoretic complexity. We find that proportion of L2 speakers is a statistically significant predictor of morphological complexity but not of information-theoretic complexity, in line with the complete cases analysis reported in Section 5.1. Another way of looking at this result is by examining the distribution of the coefficient estimate for the logit-transformed L2 speaker proportion over the $m = 100$ completed samples. This is shown in Figure 3 for morphological complexity (left) and information-theoretic complexity (right); intuitively, L2 speaker proportion ends up having an effect on morphological complexity because in that case the coefficient estimates cluster in the negative numbers, whereas in the case of information-theoretic complexity the estimates straddle zero on both sides.

Although the multiple imputation analysis agrees with the complete cases analysis in terms of the effect of L2 speaker proportion on the two kinds of complexities, we obtain divergent results for the effect of population size on information-theoretic complexity: whereas in the complete cases analysis no evidence for such an effect was found (Table 4), in the multiple imputation analysis larger population sizes are associated with higher information-theoretic complexity (Table 6).

6 Discussion

In this paper, we have provided a critical reanalysis of Koplenig's (2019) study on the effects of L2 speaker proportion and population size on the morphological and information-theoretic complexity of languages. We have argued that the categorical predictor vehicularity has a number of problems, and that a better way to probe the effect of L2 learning on language complexity is to use numerical estimates of the proportion of L2 speakers. We have also criticized the use of zero-imputation to deal with missing L2 speaker proportions, as this practice is not guaranteed to yield unbiased regression estimates.

Table 5. Fixed effect coefficient estimates for a multiply-imputed linear model with morphological complexity as dependent variable and the logit-transformed proportion of L2 speakers and log-transformed population size as independent variables, including random intercepts for language family and linguistic area (N = 1,499). FMI = fraction of missing information.

	Estimate	S.E.	t	FMI	Pr(> $\lvert t \rvert$)
(intercept)	0.730	0.036	20.21	0.224	≈ 0 ***
prop. of L2 speakers (transformed)	−0.019	0.007	−2.69	0.892	0.009 **
log(population)	−0.014	0.003	−4.41	0.263	1.201×10^{-5} ***

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

Table 6. Fixed effect coefficient estimates for a multiply-imputed linear model with information-theoretic complexity as dependent variable and the logit-transformed proportion of L2 speakers and log-transformed population size as independent variables, including random intercepts for language family and linguistic area (N = 717). FMI = fraction of missing information.

	Estimate	S.E.	t	FMI	Pr(> $\lvert t \rvert$)
(intercept)	1.298	0.048	27.04	0.021	≈ 0 ***
prop. of L2 speakers (transformed)	−0.003	0.004	−0.62	0.616	0.536
log(population)	0.022	0.003	7.87	0.062	≈ 0 ***

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

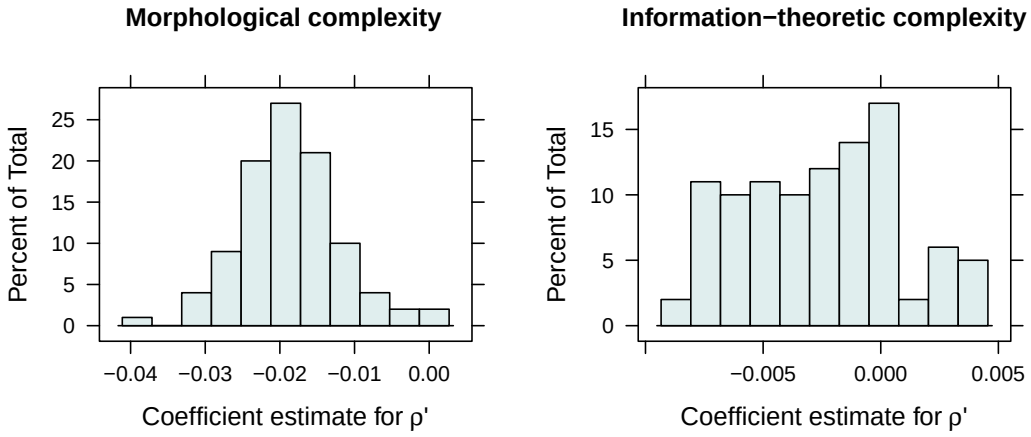

Figure 3. Distribution of the regression estimate of logit-transformed L2 speaker proportion ρ' (equation 3) over m = 100 completed copies of the data when morphological complexity is the dependent variable (left) and when information-theoretic complexity is the dependent variable (right).

We have provided two alternative analyses of the same data. In the first, complete cases analysis, missing data points are simply left out of all regressions. In the second, multiple imputation analysis, missing L2 speaker proportions are imputed using a method that both takes the attested part of the dataset into consideration and accounts for uncertainty in the resulting regression estimates that arises from the imputation process itself. Both analyses control for population size by including this predictor as a fixed effect, and for language family and linguistic area (when possible) by including these factors as random effects. Both analyses point to the same conclusion: the proportion of L2 speakers is negatively correlated with morphological complexity. In other words, populations with a higher fraction of L2 speakers tend to speak less morphologically complex languages. On the other hand, we find no evidence for an effect of L2 speaker proportion on information-theoretic complexity, on either analysis.

These findings suggest an interpretation of the role played by L2 learning in linguistic simplification that is at odds with the conclusion given in Koplenig (2019), but one that is at the same time in line with much earlier literature on the subject. We concur with Koplenig (2019) that the relationship between morphological complexity and information-theoretic complexity is unlikely to be as straightforward as suggested by Lupyan and Dale (2010), where redundancy (understood as low information-theoretic complexity) is hypothesized to be the driving force in morphological simplification. Koplenig (2019) does not find the inverse correlation between morphological and information-theoretic complexity that this causal account would predict, and in addition finds a strong and significant positive correlation between information-theoretic complexity and population size, but only a very weak correlation between information-theoretic complexity and proportion of L2 speakers. In our complete cases analysis, our results reveal no effect of L2 speaker proportion or population size on information-theoretic complexity, though this is likely related to the loss of statistical power that the complete cases analysis entails. Our multiple imputation analysis also finds no relationship between L2 speaker proportion and information-theoretic complexity, but does find an effect of population size on information-theoretic complexity, in line with the findings of Lupyan and Dale (2010) and Koplenig (2019). Clearly more research is needed into how (and whether) information-theoretic complexity relates to L2 acquisition.[7]

Importantly, our results are in line with those of Sinnemäki (2020), who shows for a sample of 66 languages that proportion of L2 speakers is a 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 (2020) observes that his results are at variance with those of Koplenig (2019), 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 both of our analyses, with and without imputation, yield similar results to those of Sinnemäki (2020) suggests that the latter assumption is likely to be the driving force in the discrepancy.

Our two reanalyses make different assumptions about the mechanism of data missingness, i.e. about why an L2 speaker proportion may not be recorded for a language in a typological database such as Ethnologue. The complete cases analysis is statistically robust only in the very specific case in which data are missing completely at random (MCAR). Thus even though this sort of conservative analysis may appear like an intuitive choice at first sight, as soon as data are missing not completely at random, the analysis cannot be trusted to yield unbiased regression estimates. For this reason, we have provided the second reanalysis, based on multiple imputation, a technique which can deal with missing at random (MAR) data.

[7] A promising start is made by Ehret and Szmrecsanyi (2019), who show that increased L2 instruction correlates with lower compressibility (i.e. higher information-theoretic complexity); they also, however, observe (i) an effect of the L1 background of their writers, and more importantly (ii) a trade-off between morphological and syntactic complexity (operationalized in terms of compressibility).

Thus multiple imputation allows us to enlarge the size of the dataset; the motivation is the same as with the zero-imputation strategy employed by Koplenig (2019), but the method can be expected to be more robust than the latter. Nevertheless, any analysis that employs imputation is necessarily worse than what could be obtained if all the data were to hand, and hence we need to also discuss the limitations of the method. There are two principal such limitations.

The first has to do with the fraction of missing data. In our case, 1,972 out of 2,143 languages in the dataset—that is to say, about 92%—are missing an L2 speaker proportion. The suggestion to impute this many missing values may seem outrageous—if more data points are missing than are attested, then how can we ever have confidence in the imputation? Indeed, a fraction of missing data of 40% has been cited as an upper bound beyond which results should only be considered hypothesis-generating (Jakobsen et al., 2017). A recent simulation study found, however, that in properly specified multiple imputation, high rates of missingness need not lead to unbiased regression estimates: according to the findings of Madley-Dowd et al. (2019), as much as 90% of missingness is tolerated by the method as long as the imputation model includes all necessary predictors. They conclude:

> A key finding of this study is that the proportion of missing data should not be used as a guide to whether to use MI (or CCA) [multiple imputation, complete cases analysis] or not—we have shown that correctly specified MI can reduce bias and improve efficiency for analysis of MAR data at any proportion of missingness (Madley-Dowd et al., 2019, 72).

In other words, as long as the data are MAR and the imputation model has been properly specified, so that all relevant predictors are included, multiple imputation can be expected to lead to analytical gains with little risk of biasing the resulting regression estimates.

This leads us to the second potential worry: the possibility that data are missing not at random (MNAR). In our case, this would mean that the probability of an L2 speaker proportion being missing from the dataset depends on that missing proportion itself. In the current state of understanding, we feel it would be premature to conclude one way or the other, but we point out that any argument to the effect that L2 speaker proportions are MNAR would need to specify a mechanism whereby such missingness arises. One possible such mechanism can be sketched along the following lines: a greater proportion of L2 speakers in a speech community increases, in general, the access that outsiders have to that community, and hence also increases the likelihood of the demographic variable of L2 speaker proportion being recorded by field typologists. However, it is also possible that such dependencies are mediated by other factors, such as population size (larger languages are more likely to be studied for demographic variables), language family (certain historically well-represented families are more likely to have these demographic variables recorded) and linguistic area (access to certain areas is, again largely for historical reasons, more likely than access to other areas). Population size and language family are taken into account by our imputation model, so we believe it does represent an analytical step forward in our understanding of the effect of L2 learning on linguistic simplification. We invite other researchers to engage critically with these findings.

Acknowledgements

The work reported here was funded by the European Research Council as part of project STARFISH (851423). Access to Ethnologue was made possible by the Communication, Information, Media Centre (KIM) of the University of Konstanz. We thank Frederik Hartmann for discussions, as well as the reviewers for their constructive criticisms and fruitful suggestions, and especially for encouraging us to pursue the missing data and imputation problem further than we had originally set out to do.

Data availability

Data and code (R, version 4.0.4; R Core Team 2021) may be obtained from https://doi.org/10.5281/zenodo.7752933.

References

Atkinson, M., Smith, K., and Kirby, S. (2018). Adult learning and language simplification. *Cognitive Science*, 42:2818–2854.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bentz, C., Ruzsics, T., Koplenig, A., and Samardžić, T. (2016). A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Berdicevskis, A. and Semenuks, A. (2022). Imperfect language learning reduces morphological overspecification: experimental evidence. *PLoS ONE*, 17:e0262876.

Dahl, O., editor (2004). *The growth and maintenance of linguistic complexity*. John Benjamins, Amsterdam.

Dryer, M. S. and Haspelmath, M. (2013). The World Atlas of Language Structures Online. https://wals.info/.

Ehret, K. and Szmrecsanyi, B. (2019). Compressing learner language: an information-theoretic measure of complexity in SLA production data. *Second Language Research*, 35:23–45.

Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press, New York, NY.

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*. Sage, Thousand Oaks, CA, 3rd edition edition.

Jakobsen, J. C., Gluud, C., Wetterslev, J., and Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Medical Research Methodology*, 17:162.

Kontoyiannis, I. (1997). The complexity and entropy of literary styles. Technical report, Stanford University, Stanford, CA. NSF Technical Report No. 97.

Koplenig, A. (2019). Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *Royal Society Open Science*, 6:181274.

Koplenig, A., Meyer, P., Wolfer, S., and Müller-Spitzer, C. (2017). The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort. *PLoS ONE*, 12:e0173614.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Lewis, M. P. and Simons, G. F. (2010). Assessing endangerment: expanding Fishman's GIDS. *Revue Roumaine de Linguistique*, 55:103–120.

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

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110:63–73.

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In Miestamo, M., Sinnemäki, K., and Karlsson, F., editors, *Language complexity: typology, contact, change*, pages 23–41. John Benjamins, Amsterdam.

Ornstein, D. S. and Weiss, B. (1993). Entropy and data compression schemes. *IEEE Transactions on Information Theory*, 39(1):78–83.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Simons, G. F. and Fennig, C. D. (2017a). *Ethnologue: Languages of the World*. SIL International, Dallas, TX, 20th edition edition. https://www.ethnologue.com/.

Simons, G. F. and Fennig, C. D. (2017b). Language status. In *Ethnologue: Languages of the World*. SIL International, Dallas, TX, 20th edition edition. https://www.ethnologue.com/.

Sinnemäki, K. (2020). Linguistic system and sociolinguistic environment as competing factors in linguistic variation: a typological approach. *Journal of Historical Sociolinguistics*, 6:20191010.

Trudgill, P. (2011). *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford University Press, Oxford.

van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press, Boca Raton, FL, second edition edition.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

Walkden, G. and Breitbarth, A. (2019). Complexity as L2-difficulty: implications for syntactic change. *Theoretical Linguistics*, 45:183–209.