Journal of Language Evolution, 2021, 99–118
doi: 10.1093/jole/lzab004
Target article

OXFORD

Crouching TIGER, hidden structure: Exploring the nature of linguistic data using TIGER values

Kaj Syrjänen [1,2,*,†], Luke Maurits [2,3,4,†], Unni Leino [1], Terhi Honkola [2,5], Jadranka Rota[6] and Outi Vesakoski[2,7]

[1]Faculty of Information Technology and Communication Sciences, Research Centre PLURAL, Tampere University, Tampere 33014, Finland, [2]Faculty of Science, Department of Biology, University of Turku, Turku 20014, Finland, [3]Faculty of Languages, Department of Linguistics and Philology, Uppsala University, Uppsala 751 26, Sweden, [4]Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig D-04103, Germany, [5]Department of Anthropology and Archaeology, University of Bristol, Bristol, BS8 1UU, UK, [6]Biological Museum, Department of Biology, Lund University, Lund, 223 62, Sweden and [7]Faculty of Humanities, Department of Finnish and Finno-Ugric languages, University of Turku, Turku 20014, Finland

*Corresponding author: kaj.syrjanen@tuni.fi

[†]Shared first authorship.

Abstract

In recent years, techniques such as Bayesian inference of phylogeny have become a standard part of the quantitative linguistic toolkit. While these tools successfully model the tree-like component of a linguistic dataset, real-world datasets generally include a combination of tree-like and nontree-like signals. Alongside developing techniques for modeling nontree-like data, an important requirement for future quantitative work is to build a principled understanding of this structural complexity of linguistic datasets. Some techniques exist for exploring the general structure of a linguistic dataset, such as NeighborNets, δ scores, and Q-residuals; however, these methods are not without limitations or drawbacks. In general, the question of what kinds of historical structure a linguistic dataset can contain and how these might be detected or measured remains critically underexplored from an objective, quantitative perspective. In this article, we propose TIGER values, a metric that estimates the internal consistency of a genetic dataset, as an additional metric for assessing how tree-like a linguistic dataset is. We use TIGER values to explore simulated language data ranging from very tree-like to completely unstructured, and also use them to analyze a cognate-coded basic vocabulary dataset of Uralic languages. As a point of comparison for the TIGER values, we also explore the same data using δ scores, Q-residuals, and NeighborNets. Our results suggest that TIGER values are capable of both ranking tree-like datasets according to their degree of treelikeness, as well as distinguishing datasets with tree-like structure from datasets with a nontree-like structure. Consequently, we argue that TIGER values serve as a useful metric for measuring the historical heterogeneity of datasets. Our results also highlight the complexities in measuring treelikeness from linguistic data, and how the metrics approach this question from different perspectives.

Key words: language evolution; TIGER algorithm; Uralic languages; simulated language data; quantitative linguistics

1. Introduction

In recent years, interest in the development and use of quantitative methods in many linguistic domains, including historical linguistics, typology, and dialectology, has increased markedly. Among the most prominent quantitative methods are those based on explicit statistical models adopted from evolutionary biology, which are used, for instance, for estimating the shapes of language families from the perspective of lexicon and typology (e.g. Dunn et al. 2008; Bouckaert et al. 2012; Honkola et al. 2013; Syrjänen et al. 2013; Chang et al. 2015; Dunn 2015; Greenhill et al. 2020), as well as exploring closely related language groups (Bowern 2012; Honkola et al. 2019) or intralingual variation (Prokić and Nerbonne 2013; Syrjänen et al. 2016; Honkola et al. 2018). While these evolutionary quantitative methods are quite capable of describing linguistic variation from a statistical standpoint, they are not by any means perfect. One key concern with the statistical approach is whether the models adequately account for languages having complex and unique histories both externally and internally, as well as whether data vary in quality. This has also been among the factors that have posed challenges for the field of quantitative historical linguistics since it first gained prominence in the 1950s (see e.g. Embleton 1986; McMahon and McMahon 2005).

Complexity in linguistic data is the result of how languages change across time. In general terms, the bulk of linguistic features are carried over from one generation of speakers to the next through language acquisition. In practice, this resembles the biological process of descent with modification, as slight changes are introduced to the linguistic variants along the way. However, language material is also transferred laterally, both between languages and within the level of intralingual populations, through borrowing and diffusion (Jacques and List 2019). In addition to this, continuously changing communicative needs of speakers can induce changes to existing linguistic features, such as e.g. semantic shifting of existing words and the introduction of new words. Alongside the aforementioned processes that shape linguistic history, languages undergo a gradual process of divergence, occurring as dialects become communicatively isolated from one another. The divergence event itself is not necessarily unambiguous; it can be affected by processes arising from population-level variation, such as so-called 'incomplete lineage sorting' (Jacques and List 2019). The result is a family of languages with historically connected features shaped by a rich variety of processes, some of which carry information that reflects descent from a common ancestor (vertical inheritance), while others reflect a history of diffusion or language contact (different kinds of horizontal transfer). This complexity is also not connected to any particular portion of language; for instance, they are as apparent in the lexicon as they are in structural or typological data (e.g. Dunn et al. 2008). Complexity in linguistic datasets can also arise from large-scale historical factors that are not directly linguistic, such as age differences between language groups, affecting the extent to which languages within a group have diverged from one another, and language extinction events, which can unpredictably affect the overall distribution of linguistic data. In addition, data quality and data quantity are also a nontrivial source for complexity in linguistic data (Wichmann et al. 2011), affecting, e.g. the amount of undetected borrowings and coding errors, as well as the overall robustness of any analysis.

While tree-like descriptions of language history form the bulk of current quantitative historical linguistic work, no unanimous opinion among linguists exists on how realistic the tree-like model is (e.g. Croft 2000; Gray et al. 2010; François 2014; Kalyan et al. 2019), due to the complexity of linguistic history and data. As it is most likely that language families are not globally homogeneous in terms of how tree-like or nontree-like they are, treating a linguistic dataset as a homogeneous package of either horizontal or vertical type of linguistic history will undoubtedly result in neglecting significant parts of the actual history. Noteworthily, in biology, which nowadays shares many quantitative techniques with linguistics, the question of whether evolutionary history is dominated by vertical or horizontal transfer of genetic material has also been actively discussed during the last decade (Doolittle and Bapteste 2007; Bapteste et al. 2009). Horizontal genetic transfer complicates in particular the evolutionary tree models of fungi, prokaryotes, and viruses (e.g. Marcet-Houben and Gabaldón 2009). Approaches for handling this complexity have also been developed, including one which operates by identifying the parts of the genome that are primarily the result of vertical processes and those that contain a primarily horizontal component (Koonin et al. 2009; Puigbò et al. 2009).

Although the question of whether linguistic change and also other types of evolution are predominantly tree-like or not is actively discussed, the toolkit for quantitatively answering such questions remains relatively small. Within the linguistic field, perhaps the two most extensive examinations of this matter are Gray et al. (2010) and Wichmann et al. (2011), both of which

focus on three techniques with which the treelikeness of datasets can be measured: NeighborNets, δ (delta) scores, and Q-residuals (see below). Other noteworthy studies on this include Nelson-Sathi et al. (2010), where minimal lateral networks were used to visualize incompatibilities in a reference tree. Also, in a recent article, Verkerk (2019) applied a Bayesian technique called the 'multiple topologies method' to explore nontree-like language history using material from four language families.

In this article, we explore a new technique for quantitatively assessing treelikeness in linguistic data. This technique is called 'Tree Independent Generation of Evolutionary Rates' or 'TIGER' (Cummins and McInerney 2011). It produces the so-called TIGER values, which quantify 'similarity in the pattern of character-state distributions' (Cummins and McInerney 2011). TIGER values were originally developed for the exclusion of parts of phylogenetic datasets that evolve too fast to retain a phylogenetic signal. They have also been used for partitioning of phylogenetic datasets, which in turn enables the researcher to account for heterogeneity better, by e.g. specifying different generalized time-reversible submodels for different parts of the partitioned dataset (see e.g. Kainer and Lanfear 2015; Rota et al. 2018). However, despite being especially popular as a proxy for evolutionary rate, we argue that TIGER values, which are a measure of similarity rather than time, can also be useful for data exploration in other ways. Here, we explore whether they could be used to assess heterogeneity of linguistic data in the form of treelikeness.

The main focus of this article is to explore the applicability of TIGER values for linguistic material using both simulated and real-world data. Simulated data were used especially for validating TIGER values as an actually working metric for measuring treelikeness, as they allow us to mimic datasets created via different kinds of historical processes, including tree-like and nontree-like divergence. In addition, simulated data also allow us to produce datasets with different degrees of treelikeness by introducing borrowings to an otherwise tree-like data.

Here, we explore how TIGER values assess four types of simulated data: tree-like data, tree-like data with borrowing, data that approximates a dialect chain, and unstructured data. Following validations with simulated data, we apply TIGER to real-world language data from UraLex 1.0, a recently released dataset of twenty-six attested Uralic languages, with cognate coding information covering 313 meanings (Syrjänen et al. 2018). UraLex is an expanded and edited version of the data used in earlier phylogenetic work on Uralic languages (Honkola et al. 2013; Syrjänen et al. 2013; Lehtinen et al. 2014).

We also compare TIGER values with two other metrics used to quantify treelikeness in linguistic data—δ scores and Q-residuals—and examine the ability of TIGER values to differentiate between different kinds of linguistic data, by comparing the TIGER values of different types of meanings from UraLex. It has been suggested that semantic categories may change at different rates (see e.g. Pagel et al. 2007; Vejdemo and Hörberg 2016; Greenhill et al. 2017). Pagel et al. (2007), for instance, have suggested that semantic categories change at different rates, which also suggests that they may have distinct evolutionary trajectories. Thus, we assessed whether TIGER values varied between the semantic categories specified by WOLD and (Tadmor 2009), which correspond with various word classes. Also, meanings from basic vocabulary lists, such as the Swadesh lists and the Leipzig–Jakarta list, are assumed to cover meanings with a more tree-like signal than the remaining 'nonbasic' vocabulary meanings in our data. We thus compared the TIGER values of meanings from standardized basic vocabulary (i.e. items from Swadesh100, Swadesh200, and Leipzig–Jakarta lists) with TIGER values of nonbasic vocabulary meanings (i.e. items from the WOLD401–500 list, introduced in Lehtinen et al. 2014). With these sanity checks, we aim to produce a reliable assessment of TIGER values as a linguistic metric.

The article is structured as follows: we begin with the Materials and methods section, where we introduce the quantitative techniques used for estimating treelikeness relevant to this article: δ scores, NeighborNets, Q-residuals, and TIGER rates. Following this, we describe both the real-world data, UraLex, as well as the simulated data used in this study, including how the generative models used to produce the simulated data work. The results begin with the validation of TIGER values as a metric for treelikeness by exploring how they perform when analyzing four kinds of simulated datasets: tree-like data, tree-like data with borrowings, dialect chain-like data, and unstructured data. The same data are also analyzed with two other metrics of treelikeness, δ scores and Q-residuals, as well as visualized with NeighborNets. Finally, we explore our real-world data from UraLex through the lens of TIGER values, exploring how this metric ranks its different meanings, semantic categories that correspond with word classes, as well as basic vocabulary versus nonbasic vocabulary subsets of the data. Finally, the main results of the article are summarized in the discussion section.

2. Materials and methods

2.1 Existing techniques for measuring treelikeness

Within the field of linguistics, perhaps the two most extensive examinations of quantitative methods for assessing whether a dataset contains a tree-like structure are Gray et al. (2010) and Wichmann et al. (2011). Both focus on three techniques: NeighborNets, δ scores and Q-residuals. All three techniques are also incorporated as part of the SplitsTree software package (Huson and Bryant 2006), a tool that is often used for preliminary exploratory analyses of basic vocabulary datasets; δ scores and Q-residuals are also available separately, for instance, as part of the Python package *phylogemetric* (Greenhill 2016).

The first technique discussed in Gray et al. (2010) and Wichmann et al. (2011), the NeighborNet, is a distance-based phylogenetic network graph (Bryant and Moulton 2004), which allows one to visually assess how tree-like a given dataset is (see Fig. 4 for NeighborNets produced in this article). A pure tree is visualized by NeighborNet similar to an unrooted tree, while nontree-like signal changes the shape of the NeighborNet; for a structured dataset such as one with an underlying tree, this is displayed as wider, partially merged branches that form a network of conflicting connections. The applicability of NeighborNets is critically discussed in, e.g. Morrison (2010), Wichmann et al. (2011), and Murawaki (2015). As Wichmann et al. (2011) note, the main disadvantage of a NeighbourNet is that it is a visual rather than a quantifiable metric. In addition to visual interpretation being generally quite subjective, interpreting NeighborNets for more complex datasets becomes increasingly challenging due to the large amount of information within each graph. Challenges such as these call for additional techniques for quantitatively measuring how tree-like dataset is.

Unlike NeighborNets, δ scores (Holland et al. 2002) and Q-residuals (Gray et al. 2010) are metrics designed for quantifying the degree of treelikeness in a dataset. Both are based on *quartets* (groups of four taxa), and estimate treelikeness based on deviations from the so-called *four-point condition*. Assuming we have four taxa, or languages in the case of linguistic data (A, B, C, and D), there are three ways in which they can be divided into two pairs. Each division corresponds to a pair of distances:

1. (A, B), (C, D); (|AB| + |CD|)
2. (A, C), (B, D); (|AC| + |BD|)
3. (A, D), (B, C); (|AD| + |BC|)

The four-point condition is satisfied if the two largest of the aforementioned three distances are identical with each other. In other words, assuming d_1, d_2, and d_3 are the three distances ordered from longest to shortest, the four-point condition is satisfied if $d_1 = d_2$. In this case, the four taxa can be represented perfectly as a tree. Treelikeness of a specific taxon can be estimated by averaging deviations from the four-point condition across all the quartets that the taxon in question participates in, and the treelikeness of an entire dataset can be estimated by averaging deviations across all the quartets of that dataset. The δ score estimates treelikeness using the formula d_1-d_2/d_1-d_3; the score is 0 if $d_1-d_3 = 0$. The Q-residual estimates treelikeness using the formula $(d_1-d_2)^2$, where d_1 and d_2 are distances normalized so that the average distance between the taxa is 1.

Both δ score and Q-residual use taxon-wise distances as input, and provide a similar output, with each language given a value that falls between 0 and 1, measuring the amount of deviation from the four-point condition. Values closer to 0 suggest that the data fit a tree-like structure, and the value increases with the existence of less tree-like structure. Notably, δ scores and Q-residuals operate at different scales, with δ scores generally being much higher than Q-residuals. The scores given for individual languages are often summarized by a mean value over all languages to obtain a single score for the whole dataset.

δ scores and Q-residuals have been compared with one another by Wichmann et al. (2011) using linguistic data; their results were in favor of the δ score as a metric over the Q-residual due to Q-residuals being sensitive to the length of terminal branches of trees, which in turn makes them correlate with what they call 'lexical heterogeneity' (estimated from their data using a version of Levenshtein distance called LDND) and age of language families. They see no obvious reason why reticulation should correlate with these factors (Wichmann et al. 2011). In practice, both δ scores and Q-residuals are actively used; e.g. δ scores and Q-residuals are both reported for a recent Bayesian phylogenetic analysis of the Dravidian language family (Kolipakam et al. 2018).

2.2 Tiger algorithm

Tree-Independent Generation of Evolutionary Rates, or TIGER (Cummins and McInerney 2011), is a nontree-based method to estimate similarities in the distribution of aligned phylogenetic data. The similarity estimates produced by TIGER (which we refer to as 'TIGER values') were originally used for excluding characters that change too fast to retain the phylogenetic signal from

phylogenetic data (Cummins and McInerney 2011). Since its publication, the original article describing TIGER has been cited ninety-four times (Web of Science, October 26 2020) in evolutionary biology literature, in over thirty different journals. Among the first biological studies that used TIGER for producing algorithmic partitioning schemes for phylogenetic inference to allow more fine-tuned phylogenetic analyses were, e.g. Rota and Wahlberg (2012), Rota et al. (2016), and Rota et al. (2018). Others have used it to clean up a phylogenetic dataset by filtering out characters that evolve fastest (e.g. Burki et al. 2016). While TIGER values were designed with phylogenetic alignment data such as nucleotide characters in mind, the algorithm can in fact be applied to any data that can be represented as a collection of multistate characters (e.g. as by Prasanna et al. 2020 to amino acid sites). In this study, we calculate TIGER values for a set of 313 multistate characters from language data, each of which encode the cognate relationship of a specific meaning.

To infer TIGER values for a dataset of aligned characters, each aligned character is first partitioned by grouping together taxa that have identical character states at that specific location of the alignment (see Fig. 1a for the general steps of the algorithm). After each position of the aligned data has been partitioned, 'partition agreements' are calculated for each character position. Partition agreements for a specific character position are calculated by comparing its taxon set partition with the partitions of every other character position; each partition agreement score records how many of the sets in the compared character position's partition are subset of one of the sets in the partition of the character position whose TIGER value we are calculating. After each partition agreement has been recorded, a TIGER value of a character position is calculated as the arithmetic mean of the partition agreement scores between that character position and all the other character positions. The resulting TIGER value is a number ranging between 0 and 1, with values closer to 1 indicative of more stable or consistent characters. Cummins and McInerney (2011) provide a detailed mathematical description of the algorithm.

In the case of linguistic data, each feature in a dataset partitions the languages into nonoverlapping subsets, on the basis of which languages share the same value for that feature. With data denoting historical relatedness, for example, the words denoting 'water' in e.g. Russian (вода), Czech (voda), Swedish (vatten), English (water), and German (wasser), are etymologically connected, and would consequently share the same value or multistate character state within a linguistic dataset of historical connections. The corresponding words in e.g. Spanish (agua), French (eau), and Italian (acqua) are also related to one another but not related to their aforementioned Germanic and Slavic counterparts; they would consequently be given a different value or multistate character state. Thus, for the meaning 'water', the languages are partitioned into two sets, S1 = {Russian, Czech, Swedish, English, German} and S2 = {Spanish, French, Italian}. For each feature in this type of dataset, TIGER values are calculated by comparing that feature's partition to the partitions of all other features, and measuring the agreement between those partitions (Fig. 1b). Agreement here is understood as the sets in one partition being subsets of the sets in many other partitions. For example, considering now words denoting 'mountain', we could partition our Indo-European languages into three sets based on their etymological relatedness: One for Russian (гора) and Czech (hora), which is a subset of the set S1 of 'water', another for Swedish and German (both berg), which is also a subset of S1 of 'water', and finally one for English (mountain), Spanish (montaña), French (montagne), and Italian (montagna). This last set here, however, does not reflect the same hierarchy than what we get with 'water', as it is not a subset of either S1 or S2 of 'water'; the reason for this is obviously the English word 'mountain', which does not originate from shared inheritance but rather through borrowing from Old French. With this in mind, the 'mountain' partition has an agreement score of 2/3 compared with the 'water' partition. Notably, the partition agreement score of 'water' compared with 'mountain' is not 2/3 but 1/2, as the S1 partition of 'water' is not a subset of any set in 'mountain', while the S2 partition of 'water' is a subset of the third set in 'mountain' (cf. Fig. 1c). Finally, the actual TIGER values for each character are calculated as the mean of the partition agreement scores between that character and every other character (Fig. 1d).

As a metric, TIGER values are based on inferring the internal consistency of the characters within a dataset based on its character states. This makes it notably distinct from both δ scores and Q-residuals, which are both based on measuring distances between taxa, or in the case of linguistic data, languages. Thus, the output of a TIGER analysis does not score each language, as δ scores and Q-residuals do, but rather each multistate character (e.g. a meaning within a lexical cognate dataset). Another difference between the three metrics is that δ scores and Q-residuals are minimized by tree-like data and maximized by nontree-like data, whereas the opposite is true with TIGER values. To some extent, TIGER values resemble consistency indices and retention

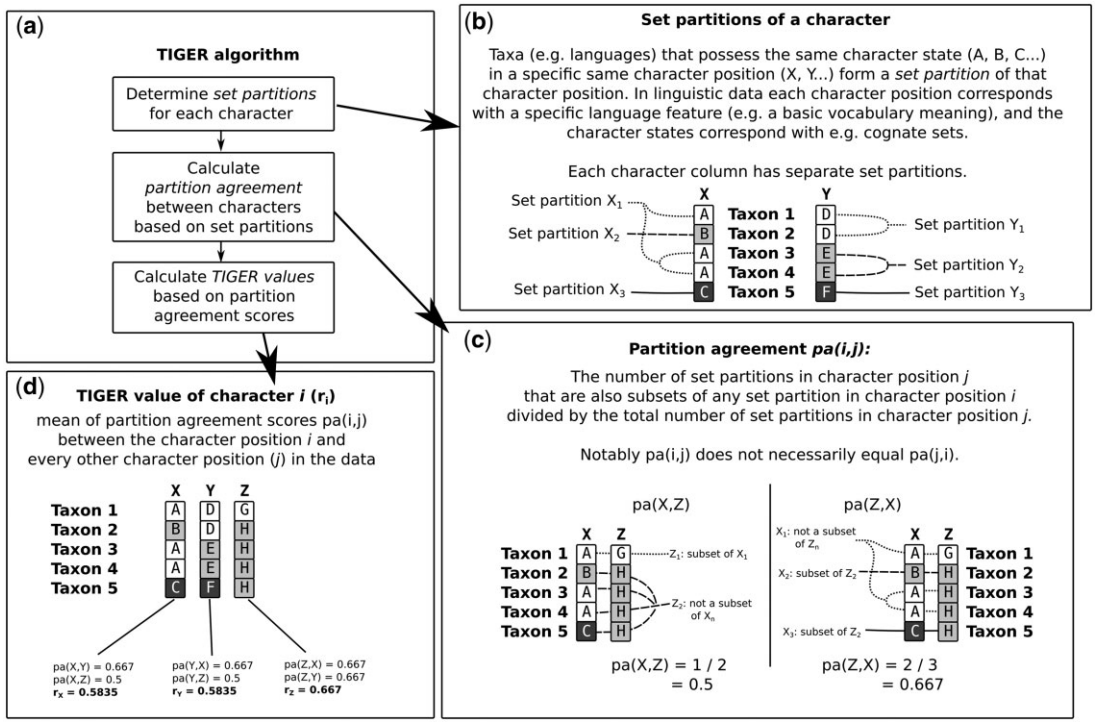

Figure 1. General overview of the TIGER algorithm. See Cummins and McInerney (2011) for a complete mathematical description.

indices, both of which measure how well phylogenetic characters fit a specific tree, as well as per-character likelihood scores (see figure 9 in Gray et al. 2010), which determine how well phylogenetic characters fit a specific model. TIGER's difference to these is that it does not score the characters against a specific tree or model but instead provides an overall measurement of the amount of hierarchical agreement across the characters in the dataset.

At least three tools are publicly available that allow the calculation of TIGER values: the original implementation by Cummins and McInerney (2011), its work-in-progress successor, and a C++ program called fast_TIGER (Frandsen et al. 2015). Unfortunately, the original TIGER tool includes a bug that causes it to miscalculate TIGER values, as stated on the tool's website. We also had trouble getting its work-in-progress successor to reliably output TIGER values. The third option, fast_TIGER, also proved unworkable for our case, as it only supports data with DNA nucleotide characters, which only have four character states (A, C, G, and T), whereas our data consist of cognate set characters, which require more character states than nucleotide data. Consequently, for the purposes of this article, we implemented a separate tool capable of calculating TIGER values for arbitrary multistate data, based on the mathematical description of the algorithm given in Cummins and McInerney (2011). A link to the tool is provided in the Supporting Material, and at the time of writing supports concurrent TIGER value calculation from three formats: FASTA files, CLDF datasets (Forkel et al. 2018), used in many linguistic phylogenetic datasets, and the CSV file format used by the simulated datasets. As a sanity check, we calculated TIGER values for random-generated character dataset resembling standard nucleotide data, which consists of states A, C, G, and T, using these different TIGER implementations. Comparisons of these analyses (see Supporting Material Table S1) showed that our calculator's TIGER values were virtually identical with the results produced by fast_TIGER, while the original TIGER implementation (v. 1.02) produced considerably different results, due to the aforementioned bug.

2.3 Establishing TIGER values as a metric of treelikeness

The first and foremost purpose of the article is to examine whether TIGER values can operate as a reliable proxy for how tree-like a dataset is. In the previous section, we described the inner workings of the TIGER

algorithm and discussed its theoretical validity as an estimator of treelikeness. However, in addition to the theoretical side, the metric also needs to be justified by experimental data. This is primarily accomplished using simulated datasets reflecting different kinds of linguistic scenarios whose parameters can be controlled. With this, we can compare how well TIGER values perform in detecting treelikeness in data, and also how they perform when subjected to data that is not based on a tree-like structure. Following this comparison, we proceed to comparing TIGER values with other metrics—δ scores and Q-residuals. The simulated tests also provide us with a point of reference when we analyze UraLex, our real-world dataset, using TIGER.

The simulated datasets presented in the main part of this article are set up so that they approximate the general size and properties of UraLex. However, we have also examined how different simulation parameters, cognate class counts, and language counts, as well as data gaps affect TIGER values; these are documented in the Supporting Material (Figs. 1–3, Supplementary Tables S2 and S3).

2.4 Comparing TIGER values, δ scores, Q-residuals, and NeighborNets

In addition to establishing how TIGER performs as a metric of treelikeness, our investigation also cross-compares TIGER values with three existing techniques designed for measuring how tree-like a dataset is: δ scores, Q-residuals, and NeigborNets. While these techniques provide a good point of comparison for how well TIGER values perform at detecting a tree-like signal, we also want to know how well especially δ scores and Q-residuals perform at this task. Consequently, δ scores, Q-residuals, and TIGER values are used to analyze both the simulated datasets as well as our real-world dataset, UraLex.

Similarly to what we do when establishing TIGER as a metric for treelikeness, we explore how each of these metrics ranks our simulated and real-world data from most tree-like to least tree-like, which we can compare with the underlying design of the simulated datasets in terms of how tree-like they should be. We also test how successful each metric is in finding the correct ranking of our simulated datasets from most tree-like to least tree-like across 100 independently generated simulations using each of the seven generative models. The results provide us with insight into how well each metric performs, but also allow us to examine how comparable these metrics actually are to one another—i.e. whether they define tree-like and nontree-like in a similar way.

While NeighborNets do not provide a quantitative assessment of the nature of the data, they serve two important functions in our investigation. First, they allow us to see how notable the differences between our simulated datasets are—i.e. whether it is possible to rank the models with respect to their treelikeness based solely on visual inspection. Second, NeighborNets also serve as a valuable means of seeing what 'nontree-like' data are for different simulated datasets.

All of the NeighborNets in this investigation are generated using *SplitsTree4* (Huson and Bryant 2006). δ scores and Q-residuals are calculated with the *phylogemetric* library (Greenhill 2016).

2.5 Language data

The real-world language data come from version 1.0 of the UraLex basic vocabulary dataset (Syrjänen et al. 2018; De Heer et al. unpublished manuscript). The dataset covers lexical reflexes (words and expressions) and historical connections ('cognate sets') for 313 meanings. Most of the meanings (226) come from standardized basic vocabulary lists: Swadesh100 (Swadesh 1955), Swadesh200 (Swadesh 1952), and Leipzig–Jakarta (Tadmor 2009). In a nutshell, basic vocabulary covers concepts that are relatively stable across time, morphologically simple, and usually resistant to being replaced by another word via, e.g. borrowing or semantic shift. This makes them useful for historical inference (see e.g. Embleton 1986; McMahon and McMahon 2005; Dellert and Buch 2018). The remaining eighty-seven meanings are from the WOLD401–500 list of 'less-basic vocabulary' (see Lehtinen et al. 2014), covering meanings that are ranked 401–500 based on the Leipzig–Jakarta list's basic vocabulary criteria and consequently fall outside of standardized basic vocabulary. We used all twenty-six attested languages as well as all 313 meanings when analyzing this dataset.

We can formulate some expectations on the nature of UraLex based on previous research. The loanword content among the lexical reflexes of UraLex has been explored in De Heer et al. (unpublished manuscript), where it is noted that there is a considerable number of loanword reflexes within UraLex—for instance, around 33% in North Saami and 16% in Komi-Zyrian (De Heer et al. unpublished manuscript). Also, earlier quantitative work with UraLex (Syrjänen et al. 2013; Honkola et al. 2013) has argued that the basic vocabulary portion of the data structures in a strongly tree-like manner, while the nonbasic vocabulary portion also contains a nontree-like signal (Lehtinen et al. 2014). Based on these results, UraLex should generally be

expected to deviate from a purely tree-like signal to some extent but not so much as to be entirely nontree-like.

UraLex is not a dataset with only one lexical reflex per meaning for a language; it has many cases where several lexical reflexes are recorded for the same meaning within a language, which are each tied to a different cognate set. These are generally found in cases where it has been difficult to define a single reflex as the representative reflex for that meaning due to, e.g. use context limitations or lack of information related to frequency of usage. However, for TIGER values we need to choose one lexical reflex (and thus, one cognate set) to represent that meaning for the language in question. In this article, we use a computational 'minimizing' strategy, where the lexical reflexes are chosen so that the total number of cognate sets for each meaning is as low as possible. This kind of strategy favors chronologically deep relationships (old historical connections between many languages) over more shallow relationships (younger connections between few languages). The Supporting Material provides a brief investigation into the effects of different cognate selection strategies for resolving synonyms on TIGER values, with the conclusion that the impact is generally minimal.

2.6 Simulated language data

When assessing the suitability of TIGER values as a measurement of tree-like historical signal in language data, it is important to consider its susceptibility to both false-negative and false-positive results. In other words, it must be established that datasets which are known to contain a tree-like signal are reliably assigned 'high scores' and also that datasets which are known *not* to contain tree-like signal are reliably assigned 'low scores'. It is also important to consider the sensitivity of TIGER values. Can they distinguish clearly between small, moderate, and large quantities of nontree-like signal, or do they offer only a coarse distinction into 'mostly vertical history' and 'mostly nonvertical history' categories?

To answer these questions, it is necessary to be able to create datasets where such details about the underlying history are explicitly known and can be directly controlled. Probabilistic generative models, such as the ones applied by Murawaki (2015), are an ideal tool for this. Different models can attempt to capture different kinds of linguistic dynamics, both tree-like and nontree-like, and large numbers of datasets can be simulated using identical model parameters, to establish the most typical properties of each model. We use generative models to create seven simulated datasets resembling language data, each with slightly different properties, and use them for testing the sensitivity of TIGER values and the other metrics in distinguishing between these different types of data, as well as their susceptibility to false negatives or positives. Different models also provide points of reference against which the TIGER values for the UraLex dataset can be compared. To facilitate this comparison, we generate each simulated dataset with the same number of languages and meanings as UraLex. There are altogether four types of simulated data that are used here: pure tree-like data, tree-like data with borrowings, dialect chain-like data, and data without any coherent internal structure. They are described in greater detail below.

2.6.1 *Purely tree-like data.*

Simulated data that contain a strong tree-like phylogenetic signal were generated by probabilistically evolving cognacy data on a randomly generated phylogenetic tree. First, a random tree with the desired number of languages was grown according to a Yule 'pure birth' model, in which the times between binary splitting events are drawn from an exponential probability distribution with a fixed rate. A unique rate value is drawn for each meaning from a gamma distribution. For each meaning, the language at the root of the tree is first assigned a cognate class. Along each branch of the tree, a number of cognate replacement events are sampled from a Poisson distribution, based on the rate value of the meaning, branch length, and cognate birth rate. If the number of replacement events is nonzero, a new cognate class is assigned to the branch's descendent node. This process continues until each leaf node has received a cognate class. Any individual cognate class is born only at a single point on the tree, and replaced cognates are never recovered. This corresponds to a 'Dollo' style evolutionary process, where once lost traits are not restored in their original form. This provides a good match for 'idealized' lexical cognate data without borrowings. The cognate birth rate can be tuned to get the desired range of extant cognate class counts and sizes.

2.6.2 *Borrowing simulation.*

For datasets generated on a tree as described above, an amount of nonvertical signal can be introduced with a coarse approximation of borrowing, including both deep and shallow borrowing events. This process is parameterized by a 'borrowing rate' between 0 and 1. First, purely vertical data are generated on a tree as described above. Following this, a second rate value is drawn from a gamma distribution, representing the borrowing

susceptibility of a meaning. Then, each node of the tree—representing the intermediary proto-stages of the simulated language family—is revisited. At each node where the node itself or its child nodes do not already serve as a borrowing source (essentially making their history 'fixed'), a Bernoulli random trial is performed to determine whether a borrowing event takes place or not; the probability of success for the trial is based on the borrowing susceptibility of the meaning, borrowing rate and branch length. In the event of a success, an attempt is made to randomly choose a borrowing source node from the same time frame as the borrower. Provided that a suitable borrowing source node is available, the cognate class of that borrowing source node overwrites the cognate class of the borrower node. Following this, the subtree starting from the borrower node is re-evolved in the same way as when generating purely vertical data, essentially simulating vertical evolution following each borrowing event. Each node of the tree is processed similarly from the root to the tips.

By varying the borrowing rate, it is possible to generate datasets with more or less nonvertical history imposed on top of the underlying tree-like evolution. This is essential for determining how sensitive TIGER values as well as the other metrics are to minute deviations from a tree-like signal. Our investigations include four simulated tree-like datasets with borrowing at different levels: trees with borrowing rates of 0.05, 0.10, 0.15, and 0.20, corresponding to an expected 5%, 10%, 15%, and 20% of datapoints at most being borrowed.

2.6.3 *Dialect chain.*

We also generated simulated data designed to loosely resemble a dialect chain. This is an important test case, as the dialect chain data have a much higher degree of internal structure than entirely unstructured data, but its predominantly spatial structure is of a notably different nature than the vertical inheritance structure that predominates the simulated datasets based on tree-structures, i.e. the purely tree-like data and the borrowing simulations. The exact process by which such data are generated is described in detail in the Supporting Material, which also includes a visual example of a small dataset (Supporting Material Fig. S5). The essential property of the resulting dataset is that there exists a one-dimensional linear ordering of the languages (approximating e.g. an East-West geographic distribution of closely related languages, or dialects within a language) such that the distribution of each cognate class along this line is consistent with the class having appeared at a single location and then spread to adjacent locations. Parameters of the generative process were set so that the number of cognate classes and their relative sizes would resemble the same properties for the UraLex data.

2.6.4 *Unstructured data ('swamp').*

The unstructured data model, affectionately termed 'swamp', generates linguistic data that lack any meaningful internal structure. For example, no two languages chosen at random are likely to be significantly more or less similar to one another than any other pair, and two languages being similar or dissimilar with regard to one feature says nothing about their likely similarity or dissimilarity with regard to any other feature. The unstructured nature of this data model puts it in stark contrast with both the tree-based data models and the spatially structured dialect chain data model. Despite being unstructured similarly, the swamp datasets are generated so that they otherwise bear a strong surface similarity to a genuine dataset. Parameters of the generative process, which is described in detail in the Supporting Material, were again set so that the number of cognate classes and their relative sizes would resemble the same properties for UraLex.

2.6.5 *Overview of the simulations.*

We produced 100 simulated datasets with the same number of meanings (313) and languages (26) as UraLex using each of the seven generative models. This provides us with 100 independent repetitions for assessing how consistently the three metrics (TIGER values, δ scores, and Q-residuals) rank the datasets in terms of treelikeness. When comparing TIGER values, δ scores, and Q-residuals in the results, we calculated mean values across all 100 repetitions. Notably, the tree-based datasets were set up so that they all used the same 100 randomly generated starting trees, but otherwise generated their data independently from one another.

2.7 Exploring the heterogeneity of linguistic categories with TIGER values

While TIGER values generally provide an overall measure for the degree of treelikeness, we also investigate the extent to which the individual values, representing a degree of structural consistency across the data, could be used for characterizing linguistic data in general. We examine the distribution of TIGER values in UraLex, determining which meanings are assigned high TIGER values and which get low TIGER values. We also compare TIGER values against the number of cognate classes per each meaning.

We further studied how TIGER values vary when divided according to two kinds of subdivision. The first kind of subdivision we explored includes WOLD's five semantic categories resembling word classes—nouns, verbs, adjectives, adverbs, and function words. Pagel et al. (2007) have suggested that semantic categories evolve at different rates; if this is true, one can hypothesize that the amount of tree-like signal measured by TIGER values should likewise vary between the classes. The second kind of subdivision was between basic vocabulary meanings—i.e. those found on Swadesh200 (Swadesh 1952), Swadesh100 (Swadesh 1955), and Leipzig–Jakarta (Tadmor 2009)—and meanings that fall outside of these lists. In the case of UraLex, these cover eighty-seven meanings from the WOLD401–500 list (Lehtinen et al. 2014), which includes meanings ranked 401–500 according to the Leipzig–Jakarta basic vocabulary criteria. Although basic vocabulary meanings versus nonbasic vocabulary meanings represent a spectrum rather than a clear-cut dichotomy, our expectation is that basic vocabulary is more tree-like than nonbasic vocabulary. In addition to exploring the distributions visually, we explored the differences between the word class categories and basic and nonbasic vocabulary also with the help of analysis of variance (ANOVA).

3. Results

3.1 Tiger values of simulated language data

We begin by considering whether TIGER values can be used to successfully distinguish between tree-like data and less tree-like data. For this, we use our seven simulated datasets. Five of the models ('pure_tree', 'borrowing_5', 'borrowing_10', 'borrowing_15', and 'borrowing_20') are trees with different degrees of borrowing, one model ('dialect') approximates a nontree-like hierarchy found in a dialect chain, and the last model ('swamp') represents nonhierarchical data.

The TIGER value distributions of the five tree-based models ('pure_tree', 'borrowing_5', 'borrowing_10', 'borrowing_15', and 'borrowing_20') are ordered in accordance with their degree of treelikeness, with 'pure_tree' having the highest mean TIGER value (0.80) and 'borrowing_20' the lowest mean TIGER value (0.73). The violin plots (Fig. 2) show that the values are fairly concentrated with all of the tree-based datasets. Furthermore, the two nontree-like datasets ('dialect' and 'swamp') are lower than any of the tree-like datasets in terms of their mean TIGER value (0.65 and 0.58, respectively). Based on this, TIGER values conform to two essential properties of a good metric of tree-like signal: first, mean TIGER values are maximized by datasets generated on trees and drop gradually as the amount of nontree-like signal increases, and second, TIGER values do not occur by chance in unstructured datasets or datasets with nontree-like structure, even when surface details of those datasets are constrained to closely match real linguistic datasets. It is worthy of note that the nontree-like datasets also stand out from the tree-like datasets based on their overall shape; the 'dialect' dataset has a somewhat wider distribution that the other datasets, whereas the 'swamp' dataset has the narrowest TIGER value distribution.

Comparing the simulated datasets with the UraLex data, we can see that this dataset shows a broad distribution of TIGER values, similar to the dialect chain simulation, but with a mean TIGER rate closer to the tree-like datasets (0.70).

In addition to the tests presented above, we also examined the sensitivity of TIGER values to datasets with fewer data points, both in the form of data with a smaller total number of meanings, and data with the same number of meanings but with more missing data points (see Supporting Material). The TIGER values inferred for all the simulated datasets were quite robust to datasets with fewer meanings than the 313 used in the tests above, with little to no effect on the general ordering of the simulations. Missing data points, on the other hand, led to an overall increase in the TIGER values, regardless of the data type; however, the general ordering of different types of data was retained here as well. Consequently, TIGER values calculated from a dataset with extensive gaps may not be directly comparable with the rates calculated from a dataset with 100% coverage, especially if the difference in coverages is considerable. The UraLex dataset, which we explored here, has fairly high data point coverage (96.2%).

We also explored what effects different parameterizations of the generative models, such as having different rates of cognate birth or different number of languages, had on the TIGER values of the simulated datasets. These are also documented in the Supporting Material. In general, it indicated that TIGER values generally performed consistently regardless of language count or cognate birth parameterization.

3.2 Tiger values compared with δ scores, Q-residuals, and NeighborNets

We next study how the performance of TIGER values compares with two existing metrics that measure a similar aspect, δ scores and Q-residuals (Fig. 3; Table S6 in Supporting Material). We also visually examine the

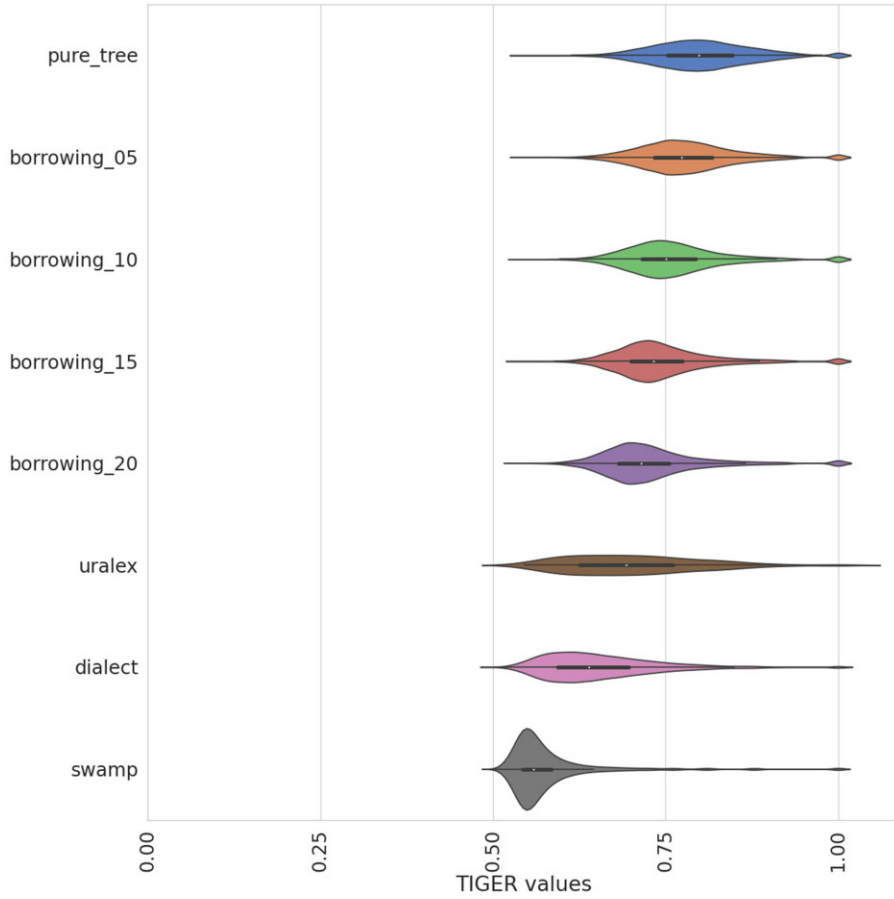

Figure 2. Violin plots of the TIGER values across seven simulated datasets and UraLex. The datasets based on a tree, from most tree-like to least tree-like, are: pure_tree, borrowing_05, borrowing_10, borrowing_15, and borrowing_20; dialect represents non-tree-like data in the form of a simulated dialect chain, and swamp represents unstructured data. The figure incorporates the TIGER value distributions from all 100 replications of each simulated dataset. The thickness of the plot reflects where the values are concentrated. The black bar in the center of the violin shows the first and third quartile, and the white point shows the median value.

characteristics of NeighborNets compared with the three metrics.

Judging by mean TIGER values, mean δ scores and mean Q-residuals from the combined results of 100 rounds of simulation (Fig. 3), all three metrics perform well in ranking the tree-based datasets from most tree-like to least tree-like. All metrics yield the expected ranking: pure tree, 5% borrowing, 10% borrowing, 15% borrowing, and 20% borrowing.

However, when examining the 100 simulated datasets from each model separately, rather than averaging over all datasets, we find that TIGER values are the most reliable of the metrics in identifying the correct order from the most tree-like dataset to the least tree-like dataset (Table 1). For example, when comparing a dataset produced with 10% borrowing on top of a tree structure to one produced with 15% borrowing, mean TIGER values rank the 10% borrowing dataset as more tree-like than the 15% dataset in 90 of the 100 runs, giving it an accuracy of 90%, whereas δ scores correctly distinguish these 76% of the time and Q-residuals 74% of the time.

While all three metrics appear to match each other quite closely in terms of how their mean values characterize the simulated datasets with some underlying tree structure, they are less consistent with respect to the simulated dialect chain and 'swamp' datasets (Figs. 2 and 3). Broadly speaking, TIGER values and δ scores behave similarly over all the simulated datasets. Both rank the swamp data as having less tree-like structure than the dialect chain data (or indeed any other dataset), followed by the dialect chain dataset as the second least

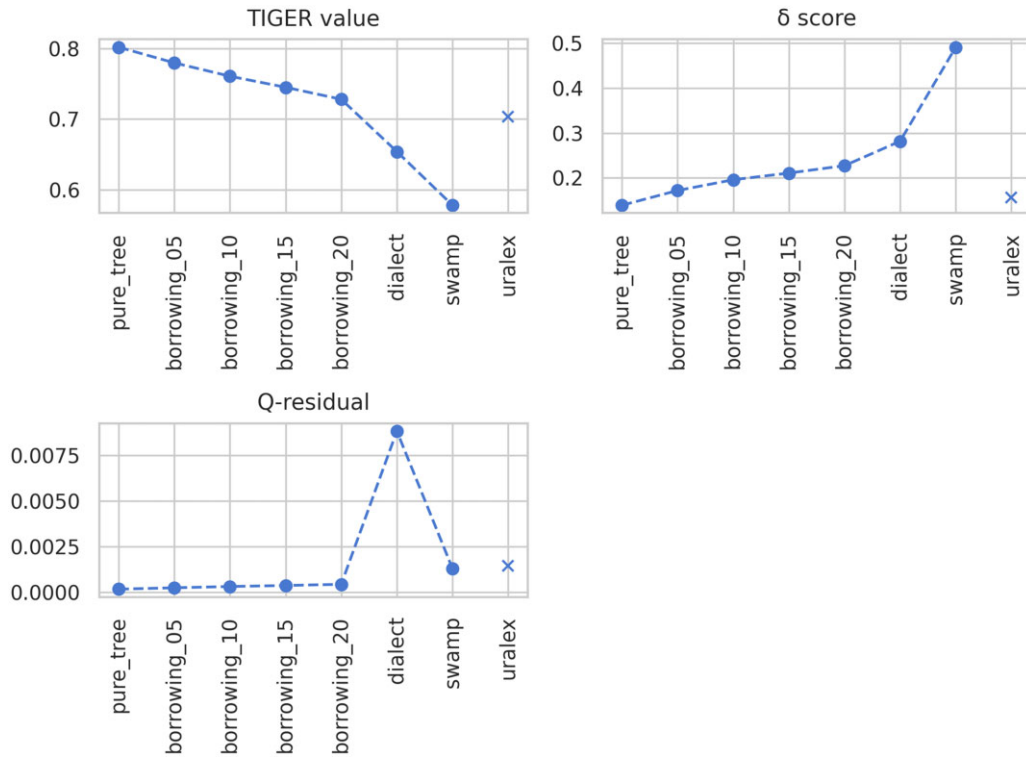

Figure 3. Mean TIGER values, δ scores and Q-residuals from the seven simulated datasets, followed by the corresponding values for the UraLex dataset. The lower value is indicative of a less tree-like signal with TIGER values, whereas the opposite is true with δ scores and Q-residuals.

Table 1. How many times a more tree-like generative model was deemed to be more tree-like than an adjacent less tree-like model across 100 rounds of simulation, based on mean TIGER values, mean δ scores and mean Q-residuals. For each row, the highest (best) result is in bold.

Consistency of different metrics in evaluating treelikeness

More tree-like versus less tree-like	TIGER value agreements	δ score agreements	Q-residual agreements
pure_tree versus borrowing_05	**93**	92	92
borrowing_05 versus borrowing_10	**95**	87	90
borrowing_10 versus borrowing_15	**90**	76	74
borrowing_15 versus borrowing_20	**91**	83	77
borrowing_20 versus dialect	99	87	**100**
dialect versus swamp	**100**	**100**	0

tree-like dataset. In contrast, the highest Q-residual score (suggesting the least treelikeness) is given not to the unstructured 'swamp' model but rather to the dialect chain model, whose Q-residual score is much higher than that of any other dataset. Q-residuals also rank the real-world UraLex data as being less tree-like than the swamp data, which contrasts especially with δ scores but also to an extent TIGER scores. This suggests that there is in fact a qualitative difference between what Q-residuals measure compared with δ scores and TIGER values. We will return to this observation later.

The NeighborNets (Fig. 4) visibly illustrate how little visual distinction there exists between a pure tree and the ones with borrowing, highlighting the usefulness of nonvisual metrics in estimating how tree-like a dataset is. Successful ranking of the pure tree and the four trees with borrowing in the correct order of treelikeness is

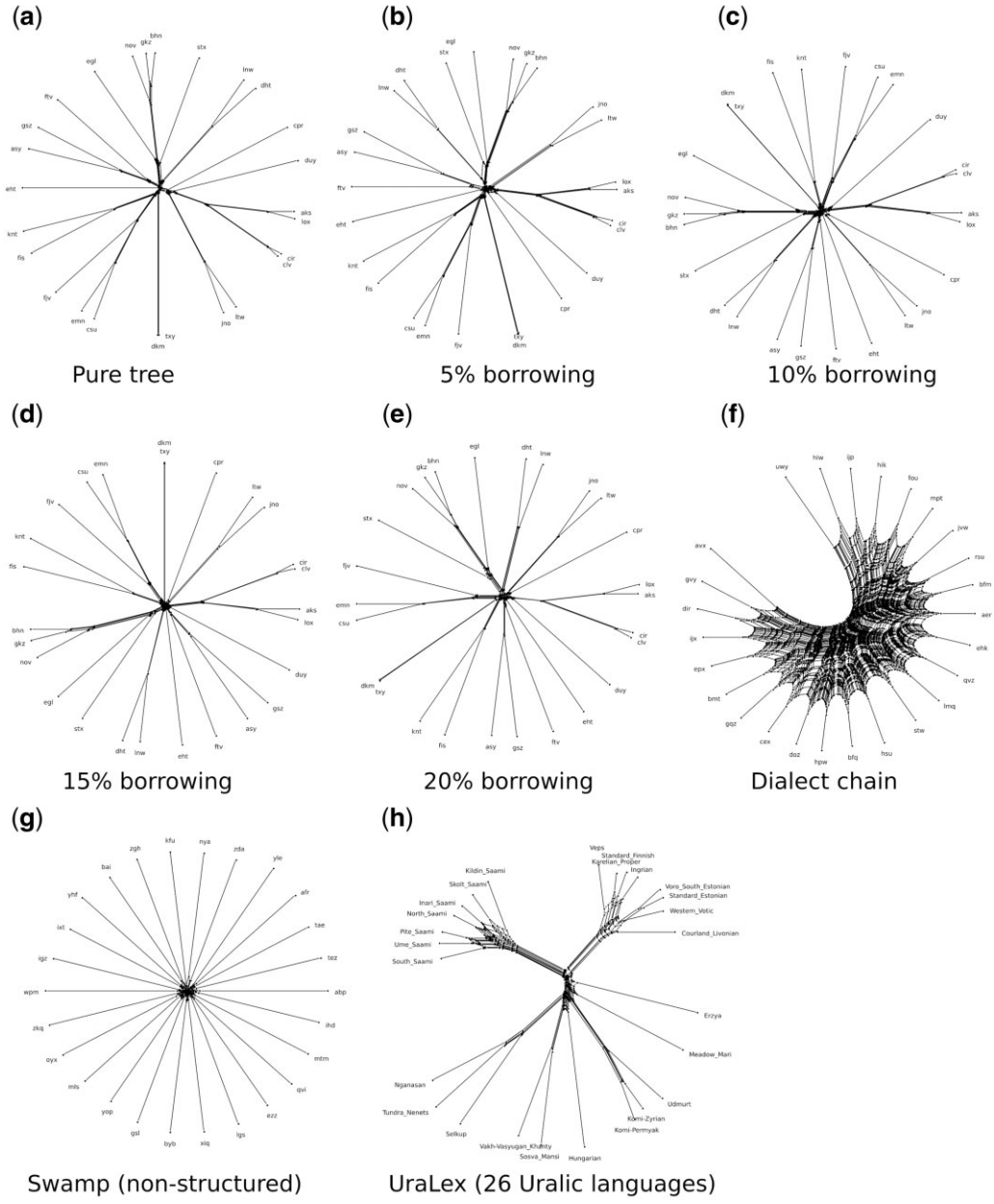

Figure 4. NeighborNets produced from the seven generative models (a–g) (in each case using the first of the 100 replications), and the UraLex lexical dataset (h). Notably, UraLex—as an uncontrolled dataset shaped by millenia of complex linguistic evolution—is visually quite distinct from the simulated datasets, each of which reflects specific linguistic scenarios in simplified forms.

difficult, if not impossible, by observing the NeighborNets alone. However, while both the dialect chain data and the 'swamp' data both lack a tree-like structure, the NeighborNets make it clear that the dialect chain data contain a large degree of nonvertical structure, leading to a 'spider web' appearance, which, in contrast, is not present in the 'swamp' data.

The datasets which display the most visible nonvertical structure in their NeighborNets, the dialect chain data and UraLex, also show the broadest spread in their distributions of TIGER values (shown in Fig. 2). At the same time, the totally unstructured swamp model data have the narrowest spread. This suggests that the spread of the TIGER value distribution, at least to a certain

extent, reflects the amount of nonvertical structure within a dataset, while the mean TIGER value tracks the amount of underlying phylogenetic signal.

3.3 Tiger values of UraLex data

Having established above that TIGER values appear to perform quite well at distinguishing data with more tree-like structure from data with less tree-like structure, we now turn our attention to what TIGER values can tell us about our real-world dataset, UraLex.

The mean TIGER value across all meanings in the UraLex dataset is 0.70 (Figs. 2 and 3). This is lower than the mean value for datasets generated according to a purely vertical evolutionary process, which should be expected for a real-world dataset which has been produced from the interaction of multiple, complicated historical processes. From the perspective of the simulated datasets, mean TIGER values would rank UraLex between a tree-like dataset with 20% borrowing rate and a dialect chain dataset; at the same time, its TIGER value distribution is quite broad, indicative of either properties similar to the dialect chain model or some other nontree-like model, or different degrees of borrowing in different parts of the data (Figs 2, 3, and 4h). Comparing the way TIGER rates characterize the dataset with the other metrics, mean δ scores position UraLex between the pure tree dataset and the 5% borrowing dataset (Figs 2 and 3), essentially establishing the dataset as a highly tree-like data. In contrast, mean Q-residuals position UraLex between the swamp dataset and the dialect chain dataset, establishing it as a highly nontree-like data.

UraLex's NeighborNet shows that this data have more visible reticulation than any of the tree-like simulated datasets, but not nearly as much as the dialect dataset. Q-residuals appear to reflect the degree of reticulation visualized by NeighborNets most prominently of the three metrics. Notably, unlike the dialect chain dataset's NeighborNet, which is considerably reticulated across the board with little visible internal hierarchy, UraLex's NeighborNet retains a mostly tree-like structure, with prominent reticulation concentrated at specific points, such as the Finnic and Saamic branches, which are both historically and geographically close to each other and also have an extensive history of language contact to the extent that they are linguistically considered to be dialect chains. There is also reticulation near the center of the UraLex NeighborNet, potentially reflecting a rapidly occurring early divergence (see e.g. Lehtinen et al. 2014). In addition to the visible similarities in their NeighbourNets with respect to the amount of reticulation, Fig. 2 indicates that the dialect chain and UraLex datasets show the broadest interquartile ranges of TIGER values out of all the datasets, suggesting a broad range of TIGER values within these datasets. This wide spread in the TIGER values is potentially characteristic of datasets with dialect-like nonvertical structure.

Taken together, the observations of UraLex could be interpreted as suggesting two kinds of structure simultaneously: a nonvertical historical structure on top of a tree-like backbone of vertical inheritance. δ scores register a prominent tree-like structure, while in contrast, the Q-residuals appear to register a prominent nonvertical structure, whereas TIGER values apparently propose a middle-ground between the two.

3.4 Tiger value distribution across meanings in UraLex data

A unique property of TIGER values, compared with alternatives like δ scores or Q-residuals, is that they provide a measurement for each character in the dataset, e.g. for each meaning when working with basic vocabulary data such as UraLex. Consequently, in addition to measuring how tree-like a dataset is as a whole, TIGER values could potentially serve a purpose in examining the linguistic internals of a dataset, such as the susceptibility of different meanings to nonvertical transmission.

UraLex data show TIGER values ranging from 0.54 to 1.00, with a mean value of 0.70 (Fig. 5; Table S5 in Supporting Material). The meanings that rank highest in terms of TIGER value are those with only one recorded cognate set: 'eye', 'I', 'name', 'two', and 'we'. Features such as these always get a TIGER value of 1.0 since every possible partition consists of subsets of the solitary set in their partition, which contains all twenty-six languages (see TIGER algorithm description). Notably, while these are undoubtedly the most unambiguous characters in the data, one may argue that they are also not the most 'tree-like' characters in the sense that they do not suggest any kind of a nested structure. The next five meanings in TIGER value order are 'five', 'four', 'hand', 'three', and 'not'; each of these meanings have two recorded cognate sets. The five meanings with the lowest TIGER value, 'float', 'twist', 'narrow', 'soon', and 'shake', have a much higher number of recorded cognate sets in UraLex (between sixteen and seventeen). Meanings with lower TIGER values also tend to have more cognate sets (Fig. 6); however, cognate set counts do not represent a one-to-one match with TIGER values, which are based on estimating the internal consistency of subsets in the data.

We now investigate how some constituent parts of the UraLex dataset are characterized by TIGER values,

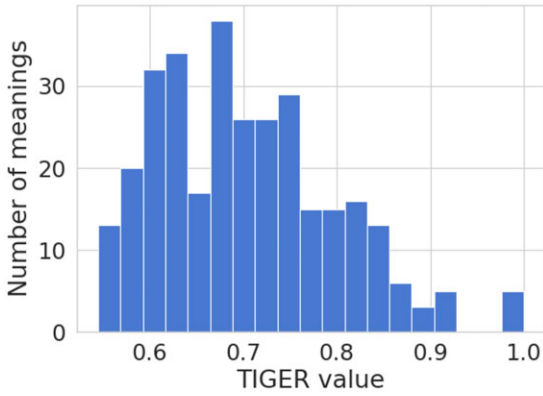

Figure 5. Histogram showing the TIGER value distribution of the meanings recorded in the UraLex basic vocabulary dataset ($n = 313$). The TIGER values for each individual meaning can be found in Table 5 in Supporting Material.

by examining what the basic vocabulary portion and the nonbasic vocabulary portion of UraLex look like from the perspective of TIGER values, and whether one can be distinguished from the other based on TIGER values. In addition to examining standardized basic vocabulary meanings, we also explore on a general level whether semantic and functional differences between meanings surface through mean TIGER values, by dividing the UraLex dataset's meanings into WOLD's five semantic categories, which correspond with word classes: nouns, verbs, adjectives, adverbs, and function words (Fig. 7).

The interquartile ranges of the TIGER values of the basic vocabulary and the nonbasic vocabulary portions of UraLex overlap considerably, suggesting that TIGER values alone do not make an unambiguous distinction between these two categories. However, the basic vocabulary portion of UraLex has a higher mean TIGER value than the nonbasic portion (0.72 for basic vocabulary versus 0.67 for nonbasic vocabulary), suggesting a more tree-like signal for the basic vocabulary portion, as indeed should be the case if basic vocabulary is to include more stable meanings. This observation was confirmed using an ANOVA analysis of log-transformed TIGER values indicated significant differences between basic and nonbasic vocabulary (see Supporting Material). However, we can note that the overall distribution of the basic vocabulary portion is also somewhat wider than that of the nonbasic vocabulary portion.

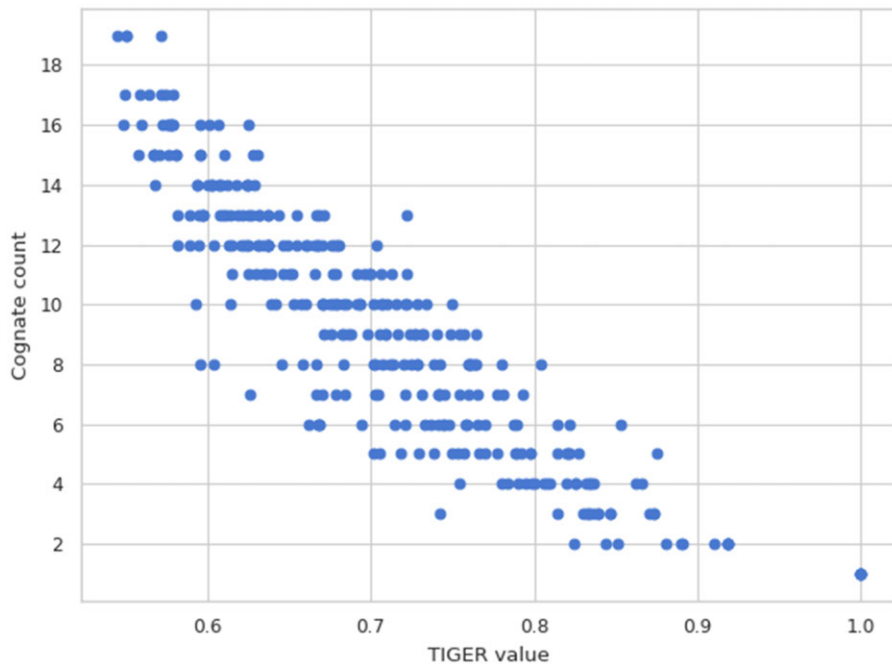

Figure 6. Scatterplot comparing TIGER values of UraLex meanings with their cognate set counts. Cognate sets have been counted after applying the 'minimizing' strategy i.e. selecting representative forms of each meaning such that the total number of cognate sets for each meaning is as low as possible. While cognate set counts and TIGER values do not match each other perfectly, they do have a strong negative correlation both when comparing the TIGER rates with UraLex's cognate set counts after applying the minimizing strategy (Pearson's coefficient: -0.90, $P < 0.0001$), as well as when comparing them with UraLex's total number of cognate sets (Pearson's coefficient: -0.84, $P < 0.0001$).

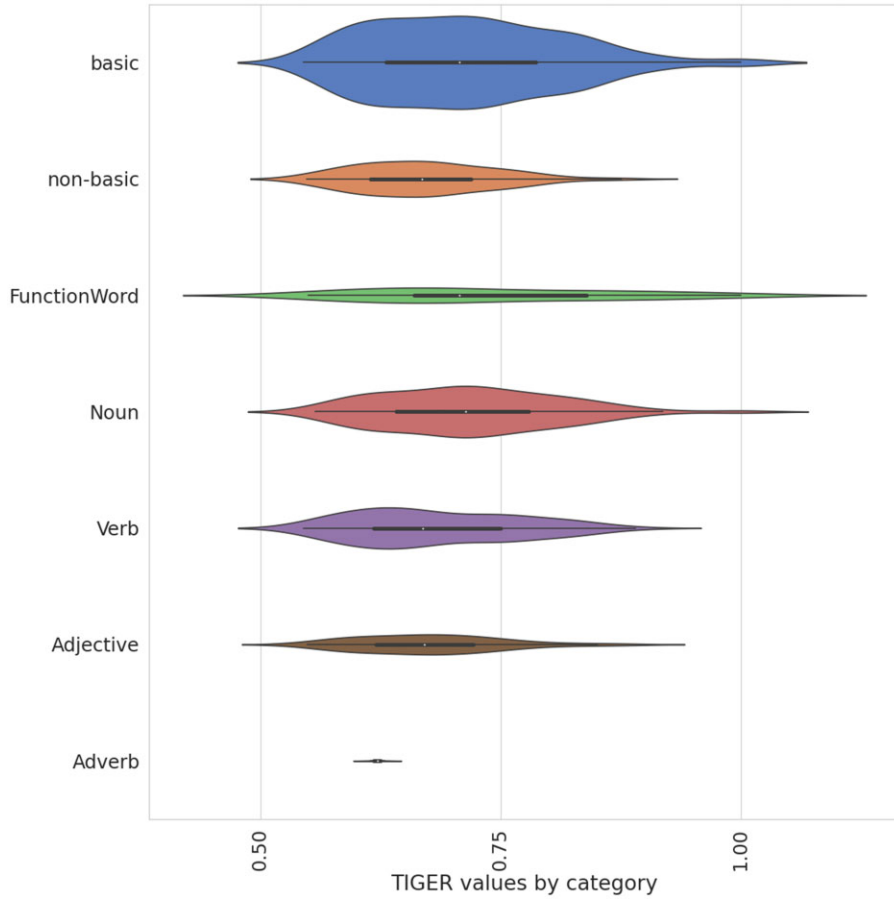

Figure 7. TIGER value distributions of UraLex basic vocabulary dataset when its meanings are subdivided into WOLD's five semantic categories corresponding with word classes: adjectives ($n = 49$), adverbs ($n = 2$), function words ($n = 37$), nouns ($n = 125$), and verbs ($n = 100$), and into basic vocabulary meanings from Swadesh200, Swadesh100 and Leipzig–Jakarta ($n = 226$) and nonbasic vocabulary meanings from the WOLD401–500 list not belonging to any of the aforementioned lists ($n = 87$). See Fig. 2 for more explanation on interpreting the plot.

Mean TIGER values would rank the five semantic classes from most tree-like to least tree-like in the following order: function words (0.745), nouns (0.718), verbs (0.684), adjectives (0.675), and adverbs (0.622; Fig. 7). Median TIGER values, on the other hand, would rank them as follows: nouns (0.714), function words (0.707), adjectives (0.671), verbs (0.669), and adverbs (0.621). As the numbers show, differences between classes are small, and there is considerable overlap in the range of TIGER values across the categories (Fig. 7). Consequently, these rankings can only be considered broad trends, and there are, e.g. some adjectives with TIGER values higher than some nouns. It should also be noted that UraLex includes only two adverbs, far less than any other semantic class. Function words ($N = 37$) have the highest mean TIGER values, but also the widest distribution. ANOVA analyses of semantic classes (excluding the underrepresented adverb class) showed nearly significant differences only between the TIGER values of nouns and verbs, and nouns and adjectives (see Supporting Material). In general, however, the semantic classes do not appear to be easily distinguished from one another on the basis of their TIGER values, indicating that none of the semantic classes appears to be considerably more 'tree-like' than the others.

4. Discussion

Based on our results, we argue that TIGER values can be used to estimate the degree of tree-like signal in language data. In fact, despite TIGER values being originally developed for another purpose, our results suggest

that they represent a notable improvement over the existing toolset for this task, which includes δ scores and Q-residuals. While all three metrics correctly ranked different data sets evolved on a phylogenetic tree from most tree-like to least tree-like when using mean values calculated over several data sets, a consideration of performance on individual datasets makes it clear that, overall, TIGER values outperform the alternatives. Beyond their improved performance, TIGER values have the additional benefit of providing information for each individual feature of a dataset. This means that TIGER values can also be used to explore the historical heterogeneity of linguistic datasets—or, indeed, of any datasets which may have been shaped by complicated cultural evolutionary processes, as TIGER values are easily computed for arbitrary multistate data.

While TIGER values can outperform δ scores and Q-residuals, it should be considered an alternative rather than a complete replacement for these metrics. Because TIGER values are calculated from multistate character data, they are incompatible with certain kinds of linguistic datasets, in particular those consisting of distance-based data. In these cases, δ scores and Q-residuals should be applied. δ scores and Q-residuals are also more straightforward to use than TIGER values for estimating the degree of reticulation of individual languages, as they are calculated across taxa rather than across aligned characters.

The effectiveness of TIGER values for the task of measuring tree-like structure can perhaps be attributed to the fact that, conceptually, their definition captures the essential nature of 'tree-like' structure quite well. The initial split in a binary phylogenetic tree partitions the languages of that tree into two sets. The subsequent splits in turn partition each of those two sets into two more, and so on. Because of this, a structure of consistently nested subsets (clades) is the defining characteristic of evolution on a tree; linguistic data which has evolved strictly via vertical inheritance on a tree will necessarily have this kind of structure, which only TIGER values directly measure.

We also did supplementary tests of how robust TIGER values were to variations in the underlying nature of the data by varying the simulation parameters and testing different language family sizes, as well as checking how data gaps affect the results (see Supplementary Material). These tests suggested that TIGER values were generally quite robust and influenced primarily by the underlying structure of the analyzed data. Notably though, the tests suggested that data gaps increase TIGER values across the board. They also showed more complex change dynamics for the TIGER values of larger language families than smaller ones. This is not generally surprising, as a larger language family is structurally more complex than a small one.

Regarding the use of TIGER values as a metric for linguistic heterogeneity, the TIGER values of various meanings in the UraLex dataset were generally distributed as one might expect based on linguistic knowledge and the nature of the UraLex dataset; in general, there exists an inverse relationship between cognate class count and TIGER value (see Fig. 6), but it is certainly not the case that simple cognate class counts provide the same kind of information as TIGER values while being simpler to compute. For instance, the meanings 'nine' and 'charcoal' both have eight cognate classes in UraLex, but 'nine' has a TIGER value of 0.76 while 'charcoal' has a TIGER value of 0.60. These differences occur because TIGER values are based on how a meaning's cognate classes are distributed across languages, which is information that cognate class counts do not encode.

We also examined how ad hoc subsets of linguistic data, rather than individual meanings or the complete dataset, were characterized by TIGER values, by subdividing the UraLex data in two different ways. The first of these was a division of the data into meanings from standardized basic vocabulary lists (Swadesh200, Swadesh100, or Leipzig–Jakarta) and nonbasic vocabulary meanings (WOLD401–500). The second was a division into five semantic categories corresponding with word classes: nouns, verbs, adjectives, adverbs, and function words. While the TIGER value distributions of basic and nonbasic vocabularies showed considerable overlap, the nonbasic vocabulary had a lower mean TIGER value of the two, suggesting a less tree-like signal. This suggests that TIGER values would behave as expected when comparing datasets with more and less stable meanings.

Semantic categories are less clearly distinguishable with TIGER values; their ranking from highest (most stable or tree-like) to lowest (least stable or tree-like) is different for mean and median TIGER values. With means, the order would be: function words, nouns, verbs, adjectives, and adverbs. With medians, the order would be: nouns, function words, adjectives, verbs, and adverbs. As a point of comparison, Pagel et al. (2007) suggested in their exploration of Indo-European rates of lexical evolution, the following order for word classes, from slowest to fastest: numbers, pronouns, special adverbs, nouns, verbs, adjectives, prepositions, and conjunctions. Notably, many of Pagel's categories, such as numbers, pronouns, prepositions, and conjunctions, could be classified as function words (Vejdemo and

Hörberg 2016), which behave differently than content words in terms of stability and borrowing-susceptibility (Vejdemo and Hörberg 2016; Thomason and Kaufman 1988). With this in mind, the rates of change suggested by Pagel et al. (2007) become somewhat more similar to our observations, with e.g. a broad distribution of TIGER values for function words versus both slow and fast function words in the results of Pagel et al. (2007), as well as nouns being more stable than verbs. However, TIGER values suggest no clear differences in treelikeness of the semantic categories. We should also point out that Pagel et al.'s inferred rates are chronological rates, specified as the number of changes per 10^4 years. A similar metric would be, for instance, the median rates of change inferred in Greenhill et al. (2017). In contrast, TIGER values, despite often regarded simply as a computationally inexpensive proxy for evolutionary rate, measure the internal consistency of a dataset, and are also not tied to an absolute time frame.

Perhaps, the most unexpected aspect of our results was how differently the three investigated metrics define tree-like data. Q-residuals ranked the nontree-based dialect chain dataset and the real-life UraLex dataset as being less tree-like than the totally unstructured dataset produced by the swamp simulation, while TIGER values and δ scores ranked UraLex as being more tree-like than both the unstructured data and the dialect chain data. Notably, both the dialect chain dataset and the UraLex dataset showed more webiness (evidence of nonvertical structure) in their NeighborNets, suggesting a possible explanation for this ranking. With this in mind, Q-residuals are apparently only an *indirect* measure of treelikeness; rather than being sensitive to the presence of tree-like structure in data (i.e. consistent nesting of subsets), they are sensitive to the absence of (at least some kinds of) nonvertical structure, and are thus maximized with datasets characterized by horizontal connections. This explains why the dialect chain data, which by design contains rich internal structure of a nontree-like variety, looks very different from all other datasets through the lens of Q-residuals. It also explains why the unstructured swamp dataset scores surprisingly well despite lacking any tree-like structure; by virtue of lacking any structure at all, it also lacks the specific kind of nonvertical structure which manifests as webiness in NeighborNets.

Despite the conceptual differences between TIGER values, δ scores and Q-residuals, all three are nonetheless valuable tools for exploring the structure of linguistic datasets. Our observations above regarding the inconsistent response of these three metrics to our datasets demonstrates an important and under-appreciated fact: language data should not be conceptualized as lying on a one-dimensional axis with tree-like data on one end and nontree-like data on the other. The results for our dialect chain and swamp models demonstrate that there is more than one way for data to be nontree-like, and that different metrics respond differently to different kinds of nontree-like structure. Furthermore, the responses of different metrics to the UraLex data, combined with that dataset's NeighborNet visualization, make it clear that a single dataset can contain multiple different kinds of structure simultaneously.

According to TIGER values, UraLex data fall near to simulated data with 20% borrowings (Fig. 3). As it happens, the actual amount of loan word present in the languages across the entire UraLex dataset is between 17 and 24% (De Heer et al. unpublished manuscript). The actual amount of borrowings was reflected neither in the δ score, which ranked UraLex between purely tree-like simulated data and simulated data with 5% borrowings, nor in Q-residuals, which scored UraLex closer to a nontree-like dataset than a tree-like one. Previous quantitative work on this data suggested that the basic vocabulary portion of the Uralic phylogeny is reasonably well resolved (high confidence values of branching events), suggesting a higher degree of treelikeness than what the results of the Q-residual analyses suggest (Syrjänen et al. 2013; Honkola et al. 2013; Lehtinen et al. 2014). However, assessing the treelikeness of different language families in light of TIGER values, as well as comparing the relationship of TIGER values and resolvedness of the phylogenetic trees are beyond the scope of the present article.

These observations only scratch the surface of a principled quantitative consideration of the structure of linguistic data. Our dialect chain model generates data which contains an internal structure which is not tree-like, but there is no reason to assume that this is the only kind of nonvertical structure which might exist in language. We believe future research should aim at combining insight from traditional historical linguistics with statistical tools for identifying latent structure in data to map out a taxonomy of structures arising from different historical processes, both vertical and nonvertical. Metrics like TIGER values and Q-residuals might end up being part of a larger toolkit of computationally inexpensive metrics (when compared with e.g. Bayesian inference of phylogeny), each one being most sensitive to a single kind of structure. The combined results of such a toolkit would then suggest an appropriate computationally intensive analysis for that specific dataset, such as Bayesian phylogenetics for clearly tree-like data, or even a collection of analyses suitable for different components of a dataset. This would hopefully lead to a

better understanding of the varied historical processes which have collectively shaped linguistic history.

Supplementary data

Supplementary data is available at *JOLEVO* online.

Acknowledgments

The authors would like to thank Mervi De Heer, Michael Dunn, Jenni Leppänen, Timo Rantanen as well as the anonymous reviewers for their valuable feedback on the manuscript.

Funding

KS, LM, TH and OV were funded by Kone Foundation projects UraLex (2013-2016) SumuraSyyni (2013-2016) and AikaSyyni (2017-2020).

References

Bapteste, E., O'Malley, M. A., Beiko, R. G., Ereshefsky, M. et al. (2009) 'Prokaryotic Evolution and the Tree of Life Are Two Different Things', *Biology Direct*, 4/34.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J. et al. (2012) 'Mapping the Origins and Expansion of the Indo-European Language Family', *Science*, 337/6097: 957–60.

Bowern, C. (2012) 'The Riddle of Tasmanian Languages', *Proceedings of the Royal Society B: Biological Sciences*, 279/1747: 4590–5.

Bryant, D. and Moulton, V. (2004) 'Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks', *Molecular Biology and Evolution*, 21/2: 255–65.

Burki, F., Kaplan, M., Tikhonenkov, D. V., Zlatogursky, V. et al. (2016) 'Untangling the Early Diversification of Eukaryotes: A Phylogenomic Study of the Evolutionary Origins of Centrohelida, Haptophyta and Cryptista.' *Proceedings of the Royal Society B: Biological Sciences*, 283/20152802.

Chang, W., Cathcart, C., Hall, D., and Garrett, A. (2015) 'Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis', *Language*, 91/1: 194–244.

Croft, W. (2000) *Explaining Language Change: An Evolutionary Approach*. Harlow, England: Longman Linguistics Library; New York, NY: Longman.

Cummins, C. A. and McInerney, J. O. (2011) 'A Method for Inferring the Rate of Evolution of Homologous Characters That Can Potentially Improve Phylogenetic Inference, Resolve Deep Divergence and Correct Systematic Biases', *Systematic Biology*, 60/6: 833–44.

Dellert, J. and Buch, A. (2018) 'A New Approach to Concept Basicness and Stability as a Window to the Robustness of Concept List Rankings', *Language Dynamics and Change*, 8/2: 157–81.

Doolittle, W. F. and Bapteste, E. (2007) 'Pattern Pluralism and the Tree of Life Hypothesis', *Proceedings of the National Academy of Sciences USA*, 104/7: 2043–9.

Dunn, M. (2015) 'Language Phylogenies'. In: Bowern Claire and Evans Bethwyn (eds) *The Routledge Handbook of Historical Linguistics*, pp. 190–211. London: Routledge.

——, Levinson, S. C., Lindström, E., Reesink, G. et al. (2008) 'Structural Phylogeny in Historical Linguistics: Methodological Explorations Applied in Island Melanesia', *Language*, 84/4: 710–59.

Embleton, S. M. 1986. *Statistics in Historical Linguistics. Quantitative Linguistics 30*. Bochum: Brockmeyer.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C. et al. (2018) 'Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics', *Scientific Data*, 5/180205.

François, A. (2014) 'Trees, Waves and Linkages: Models of Language Diversification'. In: Bowern Claire and Evans Bethwyn (eds.) *The Routledge Handbook of Historical Linguistics*, pp. 161–89. London: Routledge.

Frandsen, P. B., Calcott, B., Mayer, C., and Lanfear, R. (2015) 'Automatic Selection of Partitioning Schemes for Phylogenetic Analyses Using Iterative K-Means Clustering of Site Rates', *BMC Evolutionary Biology*, 15/13.

Gray, R. D., Bryant, D., and Greenhill, S. J. (2010) 'On the Shape and Fabric of Human History', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365/1559: 3923–33.

Greenhill, S. J. (2016) 'Phylogemetric: A Python Library for Calculating Phylogenetic Network Metrics', *The Journal of Open Source Software*, 1/2: 28.

——, Wu, C.-H., Hua, X., Dunn, M. et al. (2017) 'Evolutionary Dynamics of Language Systems', *Proceedings of the National Academy of Sciences USA*, 114/42, pp. E8822–E8829.

——, Heggarty, P., and Gray, R. D. (2020). 'Bayesian Phylolinguistics'. In: Janda, R. D., Joseph, B. D. and Vance, B. S. (eds) *The Handbook of Historical Linguistics*, Vol. 2, pp. 226–53. New Jersey: Wiley.

Holland, B. R., Huber, K. T., Dress, A. W. M., and Moulton, V. (2002) 'δ Plots: A Tool for Analyzing Phylogenetic Distance Data', *Molecular Biology and Evolution*, 19/12: 2051–59.

Honkola, T., Vesakoski, O., Korhonen, K., Lehtinen, J. et al. (2013) 'Cultural and Climatic Changes Shape the Evolutionary History of the Uralic Languages', *Journal of Evolutionary Biology*, 26/6: 1244–53.

——, Ruokolainen, K., Syrjänen, K., Leino, U.-P. et al. (2018) 'Evolution within a Language: Environmental Differences Contribute to Divergence of Dialect Groups', *BMC Evolutionary Biology*, 18/1: 1–15.

——, Santaharju, J., Syrjänen, K., and Pajusalu, K. (2019) 'Clustering Lexical Variation of Finnic Languages Based on Atlas Linguarum Fennicarum', *Linguistica Uralica*, 55/3: 161–84.

Huson, D. H. and Bryant, D. (2006) 'Application of Phylogenetic Networks in Evolutionary Studies', *Molecular Biology and Evolution*, 23/2: 254–67.

Jacques, G. and List, J.-M. (2019) 'Save the Trees: Why We Need Tree Models in Linguistic Reconstruction (and When We Should Apply Them)', *Journal of Historical Linguistics*, 9/1: 128–67.

Kainer, D. and Lanfear, R. (2015) 'The Effects of Partitioning on Phylogenetic Inference', *Molecular Biology and Evolution*, 32/6: 1611–27.

Kalyan, S., François, A., and Hammarström, H. (2019) 'Problems with, and Alternatives to, the Tree Model in Historical Linguistics', *Journal of Historical Linguistics*, 9/1: 1–8.

Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J. et al. (2018) 'A Bayesian Phylogenetic Study of the Dravidian Language Family', *Royal Society Open Science*, 5/171504.

Koonin, E. V., Wolf, Y. I., and Puigbò, P. (2009) 'The Phylogenetic Forest and the Quest for the Elusive Tree of Life', *Cold Spring Harbor Symposia on Quantitative Biology*, 74: 205–13.

Lehtinen, J., Honkola, T., Korhonen, K., Syrjänen, K. et al. (2014) 'Behind Family Trees: Secondary Connections in Uralic Language Networks', *Language Dynamics and Change*, 4/2: 189–221.

Marcet-Houben, M. and Gabaldón, T. (2009) 'The Tree versus the Forest: The Fungal Tree of Life and the Topological Diversity within the Yeast Phylome', *PLoS ONE*, 4/2: e4357.

McMahon, A. and McMahon, R. (2005). *Language Classification by Numbers*. Oxford, UK: Oxford Linguistics; New York, NY: Oxford University Press.

Morrison, D. A. (2010) 'Using Data-Display Networks for Exploratory Data Analysis in Phylogenetic Studies', *Molecular Biology and Evolution*, 27/5: 1044–57.

Murawaki, Y. (2015) 'Spatial Structure of Evolutionary Models of Dialects in Contact', *PLos One*, 10/7: e0134335.

Nelson-Sathi, S., List, J.-M., Geisler, H., Fangerau, H. et al. (2010) 'Networks Uncover Hidden Lexical Borrowing in Indo-European Language Evolution', *Proceedings of the Royal Society B: Biological Sciences*, 278/1713: 1794–803.

Pagel, M., Atkinson, Q. D., and Meade, A. (2007) 'Frequency of Word-Use Predicts Rates of Lexical Evolution throughout Indo-European History', *Nature*, 449/7163: 717–20.

Prasanna, A. N., Gerber, D., Kijpornyongpan, T., Aime, M. C. et al. (2020) 'Model Choice, Missing Data, and Taxon Sampling Impact Phylogenomic Inference of Deep Basidiomycota Relationships', *Systematic Biology*, 69/1: 17–37.

Prokić, J. and Nerbonne, J. (2013). 'Analyzing Dialects Biologically'. In: Fangerau Heiner, Geisler Hans, Halling Thorsten and Martin William (eds.) *Classification and Evolution in Biology, Linguistics and the History of Science. Concepts, Methods, Visualization*, pp. 147–61. Stuttgart: Franz Steiner Verlag.

Puigbò, P., Wolf, Y. I., and Koonin, E. V. (2009) 'Search for a 'Tree of Life' in the Thicket of the Phylogenetic Forest', *Journal of Biology*, 8/6: 59.

Rota, J., Malm, T., Chazot, N., Peña, C. et al. (2018) 'A Simple Method for Data Partitioning Based on Relative Evolutionary Rates', *PeerJ*, 6/e5498.

——, and Wahlberg, N. (2012) 'Exploration of Data Partitioning in an Eight-Gene Data Set: Phylogeny of Metalmark Moths (Lepidoptera, Choreutidae): Exploration of Data Partitioning in an Eight-Gene Data Set', *Zoologica Scripta*, 41/5: 536–46.

——, Peña, C., and Miller, S. E. (2016) 'The Importance of Long-Distance Dispersal and Establishment Events in Small Insects: Historical Biogeography of Metalmark Moths (Lepidoptera, Choreutidae)', *Journal of Biogeography*, 43/6: 1254–65.

Swadesh, M. (1952) ''Lexicostatistic Dating of Prehistoric Ethnic Contacts', *Proceedings of the American Philosophical Society*, 96: 452–63.

—— (1955) 'Towards Greater Accuracy in Lexicostatistic Dating', *International Journal of American Linguistics*, 21: 121–37.

Syrjänen, K., Honkola, T., Korhonen, K., Lehtinen, J. et al. (2013) 'Shedding More Light on Language Classification Using Basic Vocabularies and Phylogenetic Methods: A Case Study of Uralic', *Diachronica*, 30/3: 323–52.

——, Honkola, T., Lehtinen, J., Leino, A. et al. (2016) 'Applying Population Genetic Approaches within Languages: Finnish Dialects as Linguistic Populations', *Language Dynamics and Change*, 6/2: 235–83.

——, Lehtinen, J., Vesakoski, O., de Heer, M. et al. (2018) 'Lexibank/Uralex: Uralex Basic Vocabulary Dataset', *Zenodo*, https://doi.org/10.5281/zenodo.1459402. Accessed 23 March 2021.

Tadmor, U. (2009). 'Loanwords in the World's Languages: Findings and Results.' In: Haspelmath Martin and Tadmor Uri (eds.) *Loanwords in the World's Languages: A Comparative Handbook*, pp. 55–75. Berlin: Walter de Gruyter.

Thomason, S. G. and Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Vejdemo, S. and Hörberg, T. (2016) 'Semantic Factors Predict the Rate of Lexical Replacement of Content Words', *PLoS ONE*, 11/1: e0147924.

Verkerk, A. (2019) 'Detecting Non-Tree-like Signal Using Multiple Tree Topologies', *Journal of Historical Linguistics*, 9/1: 9–69.

Wichmann, S., Holman, E. W., Rama, T., and Walker, R. (2011) 'Correlates of Reticulation in Linguistic Phylogenies', *Language Dynamics and Change*, 1/2: 205–40.