Math.Comput.Sci. (2018) 12:33–50
https://doi.org/10.1007/s11786-017-0329-x

Mathematics in Computer Science

# Persistent Topology of Syntax

**Alexander Port · Iulia Gheorghita · Daniel Guth · John M. Clark · Crystal Liang · Shival Dasu · Matilde Marcolli**

Received: 15 December 2016 / Revised: 8 December 2017 / Accepted: 13 December 2017 /
Published online: 26 December 2017

**Abstract** We study the persistent homology of a data set of syntactic parameters of world languages. We show that, while homology generators behave erratically over the whole data set, non-trivial persistent homology appears when one restricts to specific language families. Different families exhibit different persistent homology. We focus on the cases of the Indo-European and the Niger–Congo families, for which we compare persistent homology over different cluster filtering values. The persistent components appear to correspond to linguistic subfamilies, while the meaning, in historical linguistic terms, of the presence of persistent generators of the first homology is more mysterious. We investigate the possible significance of the persistent first homology generator that we find in the Indo-European family. We show that it is not due to the Anglo-Norman bridge (which is a lexical, not syntactic phenomenon), but is related instead to the position of Ancient Greek and the Hellenic branch within the Indo-European phylogenetic network.

**Keywords** Linguistics · Syntax · Persistent homology

**Mathematics Subject Classification** 91F20 · 55U10 · 68P05

A. Port · I. Gheorghita · D. Guth · J. M. Clark · C. Liang · S. Dasu · M. Marcolli (✉)
Division of Physics, Mathematics and Astronomy, Caltech, 1200 E. California Blvd, Pasadena, CA 91125, USA
e-mail: matilde@caltech.edu

A. Port
e-mail: aport@caltech.edu

I. Gheorghita
e-mail: igheorgh@caltech.edu

D. Guth
e-mail: dguth@caltech.edu

J. M. Clark
e-mail: jmclark@caltech.edu

C. Liang
e-mail: clliang@caltech.edu

S. Dasu
e-mail: sdasu@caltech.edu

## 1 Introduction

This project is part of a broader ongoing investigation into the use of methods from data analysis to identify the presence of structures and relations between the syntactic parameters of the world languages, considered either globally across all languages, or within specific language families and in comparative analysis between different families.

We analyze the SSWL database [1] of syntactic structures of world languages, using methods from *topological data analysis*. After performing principal component analysis to reduce the dimensionality of the data set, in order to be able to run the Perseus software [2], we compute the persistent homology. The generators behave erratically when computed over the entire set of languages in the database. However, if restricted to specific language families, non-trivial persistent homology appears, which behaves differently for different families. By performing cluster analysis, we show that the four major language families in the database (Indo-European, Niger–Congo, Austronesian, Afro-Asiatic) exhibit different cluster structures in their syntactic parameters. This allows us to focus on specific cluster filtering values, where other non-trivial persistent homology can be found. We focus our analysis on the two largest language families covered by the SSWL database: the Niger–Congo family and the Indo-European family.

Our analysis shows that the Indo-European family has a non-trivial persistent generator of the first homology, and two persistent generators of the zeroth homology (persistent connected components), with substructures emerging at specific cluster filtering values. The Niger–Congo family, on the other hand, does not show presence of persistent first homology, and has three persistent connected components for a specific range of cluster filtering values. While the zeroth persistent homology detects clustering, the first persistent homology detects a more subtle phenomenon, namely the fact that part of the data are disposed around a nontrivial geometric structure, a closed loop.

We discuss the possible linguistic significance of persistent connected components and persistent generators of the first homology. We propose an interpretation of persistent components in terms of subfamilies, and we analyze different possible historical linguistic mechanisms that may give rise to non-trivial persistent first homology.

We focus on the non-trivial persistent first homology generator in the Indo-European family and we try to trace its origin in the structure of the phylogenetic network of Indo-European languages. The first hypothesis we consider is the possibility that the non-trivial loop in the space of syntactic parameters may be a reflection of the historical "Anglo-Norman bridge" connecting French to Middle English, hence creating a non-trivial loop between the Latin (Romance) and the Germanic subtrees. There are good linguistic reasons to exclude this hypothesis a priori, based on the fact that the Anglo-Norman bridge is known to be a lexical phenomenon and we are only analyzing syntactic data. Indeed, our data analysis of the syntactic parameters of the Romance and Germanic subtrees taken on their own demonstrates that the first persistent homology does not arises from this branch of the Indo-European family. We similarly show that it does not arise from the Indo-Iranian branch. On the other hand, the inclusion or removal of the Hellenic branch has a notable effect on both the nature of the persistent first homology and the number of persistent components. This identifies the region of the Indo-European family where the non-trivial first homology is located.

## 2 Syntactic Parameters and Data Analysis

The idea of codifying different syntactic structures through *parameters* is central to the Principles and Parameters model of syntax within Generative Linguistics, [3,4]. In this approach, each language is codified by means of a string of binary (± or $0/1$ valued) variables, the syntactic parameters, representing specific syntactic features. An excellent expository account of the parametric approach to syntax is given in [5], while a recent more technical survey on syntactic parameters is given in [6]. The comparative study of syntactic structures across different world languages plays an important role in Linguistics, see [7] for a recent extensive treatment. The study of syntactic parameters has also come to play a role in the study of Historical Linguistics and language change, see for instance [8,9].

The notion of syntactic parameters, namely the idea that syntactic structures of natural languages can be "coordinatized" by a set of binary variables, is extremely appealing from the point of view of a mathematician approaching the study of linguistic structures.

One area of criticism leveled at the Principles and Parameters model is that we lack depth of understanding of the *space* (geometric space) of syntactic parameters [10]. In particular, the model suffers from a lack of a clearly identifiable set of *independent* binary variables that can be thought of as a “universal set of coordinates”, as well as from the fact that higher-level relations between syntactic parameters have so far received insufficient attention.

From a mathematical perspective, understanding the relations between syntactic parameters corresponds to describing the geometry and topology of the space of language structures, seen as a manifold in an overall ambient space of binary variables. The first necessary step in developing such a mathematical approach to the study of syntax is investigating what topological structures can be detected in the available data of syntactic structures of world languages.

It is only in recent years, however, that accessible online databases of syntactic structures have become available. At present, publicly available data of syntactic structures can be found partly in the WALS database of [11] and mostly in the SSWL database [1], which is currently the largest available set of data on syntactic structures. Another smaller independent set of data, describing a different list of syntactic features and syntactic parameters, is collected in [9], and a more comprehensice update on this set of data is announced, but not yet published at the time of this writing.

The existence of these databases that record (even if only partially) syntactic structures across a sufficiently large number of different world languages makes syntax finally accessible to techniques of modern *data analysis*. The hope is that a computational data analysis of syntactic parameters of world languages, using various geometric and topological techniques, will elucidate potential dependencies between parameters and lead to a better understanding of the overall structure of the *geometric space of syntax*. This paper focuses on the computation of *persistent homology*, while other parts of this ongoing investigation on the “geometry of syntax” apply other mathematical methods, the results of which can be found in [12–16].

In the present work we consider only the SSWL data, since this is currently the most extensive dataset available, with 115 parameters and a set of 252 world languages. Within this database, we focus on the syntactic parameters for two of the major families of world languages: the Indo-European family and the Niger–Congo family. These are the two families that are best represented in the SSWL database, which includes 79 Indo-European languages and 49 Niger–Congo languages. Note that the Niger–Congo family is the largest language family in the world (by number of languages it comprises), although it is less extensively represented in the SSWL data than the Indo-European family. General studies of syntactic structures of Niger–Congo languages are available, see for instance [17, 18], though many of the languages within this family are still not very well mapped when it comes to their syntactic parameters in the SSWL database. The Indo-European family, on the other hand, is very extensively studied, and more of the syntactic parameters are mapped. Despite this difference, the data available in the SSWL database provide enough material for a comparative data analysis between these two families. One may worry that the lack of sufficient representation of language families other than the Indo-European in the SSWL data may impact the reliability of the analysis. It is indeed possible that some finer topological structures may exist in the distribution of syntactic features in the Niger–Congo family and that we do not see them due to the incomplete mapping of this family in the database. Since all the available resources listing syntactic features of world languages suffer from the same over-representation of the Indo-European languages, this is not a question that can be answered with certainty with the currently available data. However, one reason why we think this may not be as significant a problem as it may seem at first is the comparison with the phylogenetic trees constructed in [19] using the same SSWL syntactic data. In that setting one sees that even when restricting to small subsets of languages within a larger language family the syntactic data identify reliable phylogenetic trees, which match correctly what is known from historical linguistics.

There are other intrinsic problems with the use of the SSWL data, which one needs to keep in mind. One major issue is linguistic, namely the fact that some of the choices of binary variables recorded in the SSWL database do not reflect what linguists consider to be the “true” syntactic parameters, due to conflations of deep and surface structure. We are extending the topological analysis performed in the present paper on the SSWL data to the different set of data of [9], which better reflect the notion of syntactic parameters and is uniformly mapped for all languages in their list. We hope to return to this investigation when the announced more extensive version of their data will become available. In the meantime, we analyze the SSWL data, keeping in mind the possible shortcomings.

A notational caveat: for the purpose of this paper, we will loosely use the term "parameters" to refer to the binary syntactic variables recorded in the SSWL database as a coding of syntactic structures, even though, as mentioned above, these do not necessarily correspond to what Linguists would regard as the "true" syntactic parameters. Explicit examples of some of the binary syntactic variables recorded in the SSWL database include[1]

- *Subject Verb* this variable has the value + when in a clause with intransitive verb the order Subject Verb can be used;
- Noun Possessor this has value + when a possessor can follow the noun it modifies;
- *Initial Polar Q Marker* this has value + when a direct yes/no question is marked by a clause initial question-marker.

## 3 Persistent Homology

An important and fast developing area of data analysis, in recent years, has been the study of *high dimensional structures* in large sets of data points, via topological methods, see [20–22]. The term "high dimensional" here is used to indicate the fact that, unlike other methods of data analysis that detect cluster structures (which are topologically zero-dimensional, and correspond to just measuring the connected components of a suitable simplicial complex), topological data analysis also considers all the higher dimensional topological invariants (homology groups $H_k$ with $k > 0$) of the simplicial structures associated to the data sets. Methods of *topological data analysis* allow one to infer global features from discrete subsets of data as well as find commonalities between discrete sub-objects of given continuous objects. The techniques developed within the topological data analysis framework have found applications in fields such as pure mathematics (geometric group theory, analysis, coarse geometry) as well as in other sciences (biology, computer science) where one has to deal with large sets of data. Being qualitative in nature, topology is very well-suited in tackling these problems. In particular, topological data analysis achieves its goal by transforming the data set under study into a family of simplicial complexes, indexed by a proximity parameter. One analyzes said complexes by computing their *persistent homology*, and then encoding the persistent homology of the data set in the form of a parametrized version of a Betti number called a *barcode graph*. Such graphs exhibit explicitly the number of connected components and of higher-dimensional holes in the data. We refer the reader to [20–22] for a general overview and a detailed treatment of topological data analysis and persistent homology. An instance of this use of persistent homology is the recent study of the topology of a space of 3D images [23], where the authors determined that the barcode representation from persistent homology is isomorphic to the homology of a Klein bottle.

### 3.1 The Vietoris–Rips Complex

We recall here the general construction that associates a simplicial complex (called the Vietoris–Rips complex) to a given set of data points.

Consider a given set $X = \{x_\alpha\}$ of points in some Euclidean space $\mathbb{E}^N$. Let $d(x, y) = \|x - y\| = (\sum_{j=1}^N (x_j - y_j)^2)^{1/2}$ denote the Euclidean distance function in $\mathbb{E}^N$. The Vietoris–Rips complex $R(X, \epsilon)$ of scale $\epsilon$ over a field $\mathbb{K}$ is defined as the chain complex whose space $R_n(X, \epsilon)$ of $n$-simplices corresponds to the $\mathbb{K}$-vector space spanned by all the unordered $(n+1)$-tuples of points $\{x_{\alpha_0}, x_{\alpha_1}, \dots, x_{\alpha_n}\}$ where each pair $x_{\alpha_i}, x_{\alpha_j}$ has distance $d(x_{\alpha_i}, x_{\alpha_j}) \le \epsilon$. The boundary maps $\partial_n : R_n(X, \epsilon) \to R_{n-1}(X, \epsilon)$ with $\partial_n \circ \partial_{n+1} = 0$ are the usual ones determined by the incidence relations of $(n + 1)$ and $n$-dimensional simplices. For $n \ge 0$, one denotes by

$$\begin{aligned} H_n(X, \epsilon) &:= H_n(R(X, \epsilon), \partial) \\ &= \mathrm{Ker}\{\partial_n : R_n(X, \epsilon) \to R_{n-1}(X, \epsilon)\}/\mathrm{Range}\{\partial_{n+1} : R_{n+1}(X, \epsilon) \to R_n(X, \epsilon)\} \end{aligned}$$

[1] See http://sswl.railsplayground.net/browse/properties for a list and description of all the syntactic featured covered by the SSWL database.

the $n$-th homology with coefficients in $\mathbb{K}$ of the Vietoris–Rips complex. When the scale $\epsilon$ varies, one obtains a system of inclusion maps between the Vietoris–Rips complexes, $R(X, \epsilon_1) \hookrightarrow R(X, \epsilon_2)$, for $\epsilon_1 < \epsilon_2$. By functoriality of homology, these maps induce corresponding morphisms between the homologies, $H_n(X, \epsilon_1) \to H_n(X, \epsilon_2)$. A homology class in $H_n(X, \epsilon_2)$ that is not in the image of $H_n(X, \epsilon_1)$ is a birth; a nontrivial homology class in $H_n(X, \epsilon_1)$ that maps to the zero element of $H_n(X, \epsilon_2)$ is a death, and a nontrivial homology class in $H_n(X, \epsilon_1)$ that maps to a nontrivial homology class in $H_n(X, \epsilon_2)$ is said to persist. Mapping the deaths, births, and persistence of a set of generators of the homology as the radius $\epsilon$ grows gives rise to a barcode graph for the Betti numbers of these homology groups. Those homology generators that survive only over short intervals of $\epsilon$ radii are attributed to noise, while those that persist for longer intervals are considered to represent actual structure in the data set.

### 3.2 What Does Persistent Homology Mean?

It is in general a difficult problem in persistent topology of data to interpret what the generators of the persistent homology groups “mean” in terms of a theoretical model underlying the data. The structures we observe do not, at present, have an obvious explanation in terms of Linguistic theory and of the Principles and Parameters model of syntax.

As we will see in the rest of this paper, the zero-th persistent homology (persistent connected components), which is a way of detecting how data points cluster together, appears to match the subdivision of a language family into its main historical linguistic subfamilies. For example, one finds (in a suitable range of scales) two persistent components in the Indo-European family, corresponding to the European and the Indo-Iranic branch, and (also in a suitable range of scales) three persistent components in the Niger–Congo languages, which seem to correspond to the Mande, Atlantic–Congo, and Kordofanian subfamilies. It is more difficult to explain in historical-linguistic terms what persistent generators of the first homology correspond to. We discuss here below two different mechanisms that can give rise to an $H_1$-generators and their respective interpretation.

It is also possible that persistent generators of $H_1$ in syntactic parameters may provide information on homoplasy in syntax, which should be compared with the analysis of Warnow and collaborators, based on other types of linguistic data, [24].

Certainly, the presence of persistent homology in the syntactic parameter data, and its different behavior for different language families begs for a better understanding of the formation and persistence of topological structures from the historical-linguistics viewpoint, and from the viewpoint of Syntactic Theory. The preliminary results presented in this paper should be considered only as a starting point for such an investigation.

### 3.3 Linguistic Significance of Persistent Homology

When we analyze the persistent topology of different linguistic families (see the detailed discussion of results in § 5), we find different behaviors in the number of persistent generators in both $H_0$ and $H_1$. As typically happens in many data sets, the generators for $H_n$ with $n \geq 2$ behave too erratically to identify any meaningful structure beyond topological noise (this problem is discussed for instance in [25], although cases of topological data analysis that show a clear pattern of non-trivial persistent homology groups $H_k$ with $k \geq 2$ occur for example in [26]).

These empirical observations on the occurrence of higher persistent homology can be better understood in terms of general results on the persistent topology of random complexes (see for instance [27]). In the case of random complexes it is shown that for each $k > 0$ the persistent homology group $H_k$ has two thresholds, identified as a value of the scale $\epsilon$ expressed as a function of the homology dimension $k$, the number of points $n$, and the dimension $d$ of the ambient Euclidean space, where the groups first pass from vanishing to non-vanishing and then again from non-vanishing to vanishing. Thus, depending on the number of data points (in our case number of languages) and the ambient space dimension (in our case number of syntactic parameters) one expects to see the vanishing of certain ranges of homology groups, at least if these data sets were behaving like random complexes. The deviance from the expected random complex behavior can be seen as another indicator of the presence of structures given by syntactic relations, just as in [15] the deviance from the random codes behavior is used as an indicator of the effect of syntactic relations. We will not pursue in this paper the comparison with the random complexes behavior, but we will return to it in future work.

In general, the rank of the $n$-th homology group $H_n$ of a complex counts the "number of $(n+1)$-dimensional holes" that cannot be filled by an $(n+1)$-dimensional patch. In the topological analysis of a point cloud data set, the presence of a non-trivial generator of the $H_n$ at a given scale of the Vietoris–Rips complex implies the existence of a set of data points that is well described by an $n$-dimensional set of parameters, whose shape in the ambient space encloses an $(n+1)$-dimensional hole which is not filled by other data in the same set. In this sense, the presence of generators of persistent homology reveal the presence of structure in the data set.

In our case, the database provides a data point for each recorded world language (or for each language within a given family), and the data points live in the space of syntactic parameters. In order to render the computation of persistent homology more manageable, it is convenient to reduce the dimension of the ambient space by first performing a principal component analysis: this step will be discussed more in detail in § 4. The presence of an "$(n+1)$-dimensional hole" in the data (a generator of the persistent $H_n$) shows that (part of) the data cluster around an $n$-dimensional manifold that is not "filled in" by other data points. Possible coordinates on such $n$-dimensional structures represent relations among the syntactic parameters over certain linguistic (sub)families. It should be pointed out, however, that it is in general a difficult problem to obtain explicit coordinate functions that parameterize the $n$-dimensional structures detected by the generators for the persistent $H_n$ homology group.

Since the only persistent generators we encountered are in the $H_0$ and $H_1$, we discuss more in detail their respective meanings.

### 3.4 Linguistic Significance of Persistent $H_0$

The rank of the persistent $H_0$ counts the number of connected components of the Vietoris–Rips complex. It reveals the presence of clusters of data within which data are closer to each other (in the range of scales considered for the Vietoris–Rips complex) than to any point in any other component. Thus, a language family exhibiting more than one persistent generator of $H_0$ has linguistic parameters that naturally group together into different subfamilies. It is not known, at this stage of the analysis, whether in such cases the subsets of languages that belong to the same connected component correspond to historical linguistic subfamilies or whether they cut across them: a more detailed analysis will be needed, which we plan to carry out in future work. We will give some evidence, in the case of the Indo-European and the Niger–Congo families, in favor of matching persistent generators of the $H_0$ to major historical linguistic subfamilies within the same family. Certainly, in all cases, the connected components identified by different generators of the persistent $H_0$ can be used to define a grouping into subfamilies, whose relation to historical linguistics remains to be investigated.

It is important to keep in mind here that the concept of subfamilies of a language family has an independent definition based on historical linguistics, which is mostly based on the study of the historical development of languages at the morphological and lexical rather than the syntactic level, as well as on additional information, such as archaeological data and population genetics. Thus, the question we are considering here is whether the clustering structure determined by the persistent $H_0$ of syntactic data at different scales $\epsilon$ matches the subdivision into subfamilies determined by historical linguistics on the basis of independent data of a completely different nature.

Even when one only considers syntactic data, a different method can be adopted to obtain a subdivision into subfamilies of a linguistic family, which is based on the computational reconstruction of a phylogenetic tree, [9,14,19]. The branching structure of the tree determines the subdivision into subfamilies. It is shown in [19] that the subfamilies identified by this method match known ones determined by the historical linguistics methods described above, at least in the case of subfamilies of the Indo-European family. Thus, even by just working with syntactic data, it is a nontrivial question whether the clustering structure described by the behavior of the persistent $H_0$ at different scales matches the branching structure of the phylogenetic tree determined by the method of phylogenetic algebraic geometry described in [14,19], based on techniques in mathematical biology, [28].

### 3.5 Linguistic Significance of Persistent $H_1$

The presence of an $H_1$-generator also means that part of the data (corresponding to one of the components of the Vietoris–Rips complex) clusters around a one-dimensional closed curve. More precisely, one can identify the first homology group $H_1(X)$ of a space with the group of homotopy classes $[X, S^1]$ of (basepoint preserving) maps $f : X \to S^1$ from $X$ to the circle $S^1$. This means that, if there is a non-trivial generator of the persistent $H_1$, then there is a choice of an angle coordinate on the circle that best describes that part of the data. The freedom to deform the map by a homotopy makes it possible to seek a smoothing of the angle coordinate on the circle (a priori the curve is only topologically a circle, not smoothly). It is not obvious how to interpret these circles from the linguistic point of view. The fact that a generator of the $H_1$ represents a 2-dimensional hole means that, given the data that cluster along this circle, no further data points determine a 2-dimensional surface interpolating across the circle. As the topological structures we are investigating stem from a Vietoris–Rips complex that measures proximity between syntactic parameters of different languages, we can propose a heuristic interpretation for the presence of such circles as the case of a (sub)family of languages where each language in the subfamily has other “neighboring” languages with sufficiently similar syntactic parameters, so that one can obtain a path through the whole subfamily via changes of syntactic parameters described by a single circle coordinate, while parameter changes that move along two-dimensional manifolds and interpolate between data points on the circle cannot be performed while remaining strictly within the same (sub)family.

Two different models can be proposed to account for how a non-trivial generator of the persistent first homology can arise. They point to different possible explanations in historical-linguistic terms.

As shown in Fig. 1, the first model is a typical Hopf bifurcation picture, where a circle arises from a point (with the horizontal axis corresponding to time). This model would be compatible with a phylogenetic network of the corresponding language family with the structure of a tree, where one of the nodes generates a set of daughter nodes whose points in the parameter space are disposed in such a way that they determine a nontrivial loop in the Vietoris–Rips complex. The second possibility is of a line closing up into a circle. This may arise in the case of a language family whose phylogenetic network is not a tree and already contains a loop that closes off two previously distant branches. There are well known cases where the phylogenetic network of a language family is not necessarily best described by a tree. For example, when one considers languages at the *lexical* level, a very well known case is provided by the Anglo-Norman bridge in the network of the Indo-European languages, which creates a loop between the Romance and the Germanic languages via a historical connection between French and Middle English. However, it is important to point out that the presence of such a loop in a phylogenetic network of a language family constructed on the basis of lexical data does not imply that this loop will also exist at the syntactic

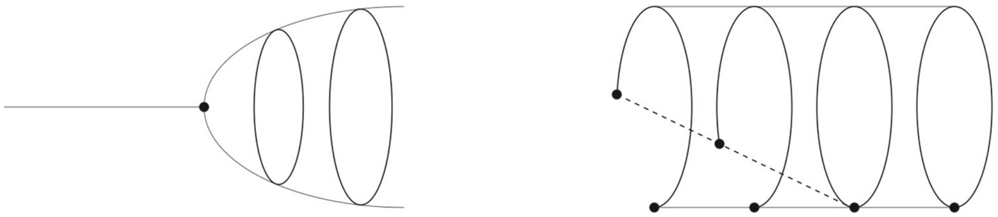

**Fig. 1** Two models of the development of a non-trivial loop in the space of parameters. The first case shows the Hopf bifurcation model, where a loop emerges from a single point, while the second case shows a bridge obtained by closing an open string into a loop. These two models are discussed in § 3.5. We argue in § 5 that the loop detected in the Indo-European language family is more likely of the first rather than the second type

level and be detectable in the form of a non-trivial first persistent homology of the set of syntactic parameters. For example, it is well known that the Anglo-Norman bridge is a lexical and not a syntactic phenomenon.

Conversely, persistent first homology alone is no guarantee that loops may be present in the phylogenetic network, even for the case of a network constructed on the basis of syntactic rather than lexical data. For example, the persistent homology loop may be due to possibilities such as the Hopf bifurcation described above. Such a picture would not necessarily introduce loops in a phylogenetic tree, as it would only represent an arrangement between various branches of a tree stemming out of a common root. Thus, one cannot infer the topology of the historical phylogenetic network directly from the presence or absence of a persistent $H_1$. The only inference that can be drawn is that a persistent $H_1$ may be due to a phylogenetic loop (in phylogenetic networks based on syntactic data, as in [9,14] for instance) as one of the possible causes. Conversely, one can read the absence of non-trivial persistent first homology as a suggestion of the fact that the phylogenetic network may indeed be a tree and that (at least purely at the level of syntax) no loops occur in the historical development of that family.

We will discuss this point more in detail in the case of the Indo-European language family. This is a very good example, which shows how the possible correlation between loops in the space of syntactic parameters and in the phylogenetic network is by no means an implication. Indeed, the Indo-European language family contains both a known loop in the phylogenetic network based on lexical data, due to the Anglo-Norman bridge and a non-trivial generator of the persistent $H_1$ in the SSWL data of syntactic structures. However, we will show using our topological data analysis method that these two loops are in fact unrelated, confirming the known linguistic fact that the Anglo-Norman bridge is only a lexical phenomena which does not leave a detectable structure on syntax. The linguistic meaning of this “circle” in the syntactic data of the Indo-European family is not immediately evident. As we discuss more in detail below, it seems to reflect certain syntactic proximities between the Hellenic branch of the Indo-European family and other branches, including some of the Slavic languages: a phenomenon that is not seen at the level of the phylogenetic trees.

## 4 Data Analysis Procedure

The SSWL database [1] is first imported into a pivot table in Excel. The on-off parameters are represented in binary, in order to compute the distances between languages as a Hamming distance. However, the parameter values are not known for many of the languages in the database: over one hundred of the languages have, at present, less than half of their parameters known. Thus, we replace empty language parameters with a value of 0.5. We obtain in this way an initial set of 252 languages, each with 115 different parameters. The effect of our choice of encoding missing parameters as ambiguous, by assigning to them a 0.5 value, is then analyzed using the completeness threshold described below, which reduces the size of the set of languages depending on the number of parameters known. This allows us to identify a range in which the results are robust and likely more reliable.

We then proceed to our analysis based on the results from Perseus homology software [2]. This is achieved through a series of Matlab scripts.[2] The script named `data_select_full.m` allows for selection of subsets of the raw data. It performs Principal Component Analysis on the raw parameter data and saves it to a text file for use in Perseus. The data are encoded geometrically through the associated Vietoris–Rips complex. This script has two important parameters: a completeness threshold, and a percent variance to preserve. The completeness threshold removes the languages below an assigned threshold of known parameters. The percent variance allows us to reduce the dimensionality of our data.

The completeness threshold is necessary, since as mentioned above several of the languages in the SSWL database are, at this stage, poorly mapped, with a significant part of the parameters missing. By performing our analysis repeatedly while varying the completeness threshold parameter, hence by only looking at the subset of languages with at least an assigned percentage of parameters known, we can identify the range in which the results are robust and measure to what extent they can be affected by the missing data.

[2] A repository of the code used for this project is available at https://github.com/cosmicomic/cs101-project5.

The dimensional reduction performed via Principal Component Analysis is required in order to make the computation of the persistent homology manageable in terms of computational time. A related interesting question, which we will return to in upcoming work, is to describe the linguistic meaning of the principal components. These can be seen as a weighted mixture of different syntactic parameters, hence they contain some potentially interesting information on syntactic relations.

The next script, named `barcode.m`, is used to create barcode graphs for data visualization. Perseus outputs the birth and death times for each persistent homology generator, which are then used to construct the barcode graph of the persistence intervals. The intervals in the barcode graph show the range of scales $\epsilon$ through which a non-trivial generator of the persistent homology continues to exist. Thus, the barcode graph visualizes the structure of the homology of the Vietoris–Rips complex at different scales and shows the persistent generators. The radii in our complexes are incremented by 1% of the mean distance between languages.

Data analysis was initially set up as a three step process: select the data with the script `data_select_full.m`, analyze it with Perseus, and use `barcode.m` to visualize the results. The final script, named `run_all`, streamlines this process under a single input command.

Finally, our analysis includes examining how many data points belong to clusters of points at any given radius. Clusters are constructed by creating $n$-spheres of uniform radius centered at each data point. If the $n$-spheres of two data points overlap, then those data points are in the same cluster. A non-trivial cluster is one containing at least two data points. The scripts `group_select.m` and `graph_clusters.m` make it possible to visualize the number of clusters and non-trivial clusters as radius increases.

## 5 The Persistent Topology of Linguistic Families

A preliminary analysis performed over the entire set of languages in the SSWL database shows that the non-trivial homology generators of $H_1$ and $H_2$ behave erratically. Moreover, there are too many generators of $H_2$ and $H_3$ to draw any meaningful conclusion about the structure of the underlying topological space.

One can see the typical behavior represented in Fig. 2. In part (a) of Fig. 2 (top-left) we included the languages with more than 60% of the parameters known, while in part (b) (top-right) we removed all languages with more than 20% of the parameters unaccounted for. Here percentage of parameters is with respect to the largest number of syntactic parameters considered in the SSWL database. One can compare this with the case of a randomly generated subset of languages, presented in the third graph of Fig. 2. Notice that, while in the cases represented in the first two graphs of Fig. 2 there is noise in the $H_1$ and $H_2$ region which prevents a clear identification of persistent generators, the homology of random subsets of the data, as displayed in the part (c) (bottom) of Fig. 2, is relatively sparse, containing only topologically trivial information. This observation leads us to the hypothesis that the behavior seen in Fig. 2 stems from a superimposition of some more precise, but non-uniform, topological information associated with the various different linguistic families.

In order to test this hypothesis, we examine specific language families as an additional method of data filtering. We choose the four largest families represented in the original database: Indo-European with 79 languages, Niger–Congo with 49, Austronesian with 18, and Afro-Asiatic with 14. Although some of the languages in the database included latitude and longitude coordinates, these are ignored when determining language families, since it is well known that geographic proximity is not an indicator of linguistic relatedness.

### 5.1 Cluster Structures in Major Language Families

A first observation, when comparing syntactic parameters of different linguistic families, is that they exhibit different cluster structures in the data of syntactic parameters. This is illustrated in Fig. 3, in the case of the four largest families in the SSWL database.

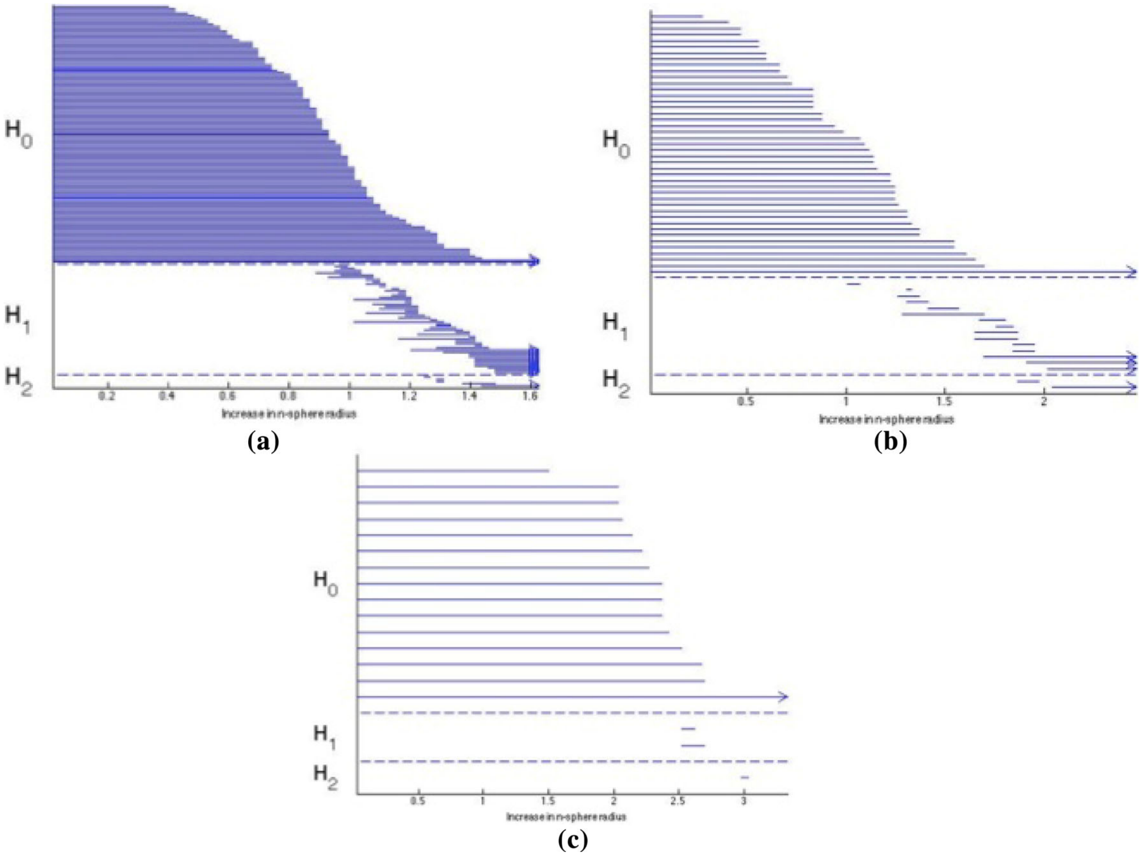

**Fig. 2** The persistent homology *barcode graphs* (defined in § 3.1) are shown for subsets of SSWL languages in the cases: **a** languages with 60% of parameters known and 60% of variance preserved; **b** with 80% of parameters known and 60% of variance preserved; **c** an example of barcode graph for a subset consisting of a random choice of 15 languages in the database with 100% of variance preserved. These examples show that persistent generators behave erratically when computed on the entire dataset (not subdivided by language families): this is discussed more in detail in § 5

We want to focus our analysis on specific cluster filtering values, which should be selected so that they correspond to regions in the cluster analysis plots where interesting structure occurs. Since the cluster structure can be significantly different for different language families, we also want to choose cluster filtering values in the overlap of the regions of interest of the different language families we wish to compare. Cluster values satisfying both of these properties are considered appropriate.

Based on this cluster analysis, we focus on the cases of the Indo-European and the Niger–Congo language families and we search for nontrivial generators of the first homology $H_1$ in appropriate ranges of cluster filtering.

The cluster analysis of Fig. 3 suggests that cluster filter values between 150 and 200 may provide additional interesting information. In particular we compute barcode diagrams corresponding to cluster filtering values 165 and 190, since at those values one sees interesting structures in the persistent homology, see Figs. 4 and 5.

## 5.2 Indexing in Barcode Graphs

In the graphs presented in the following subsections § 5.3–5.5, the barcode graphs are labeled by a set of three indices. The first two indices refer to the Principal Component Analysis and the third index to the runs of the Perseus

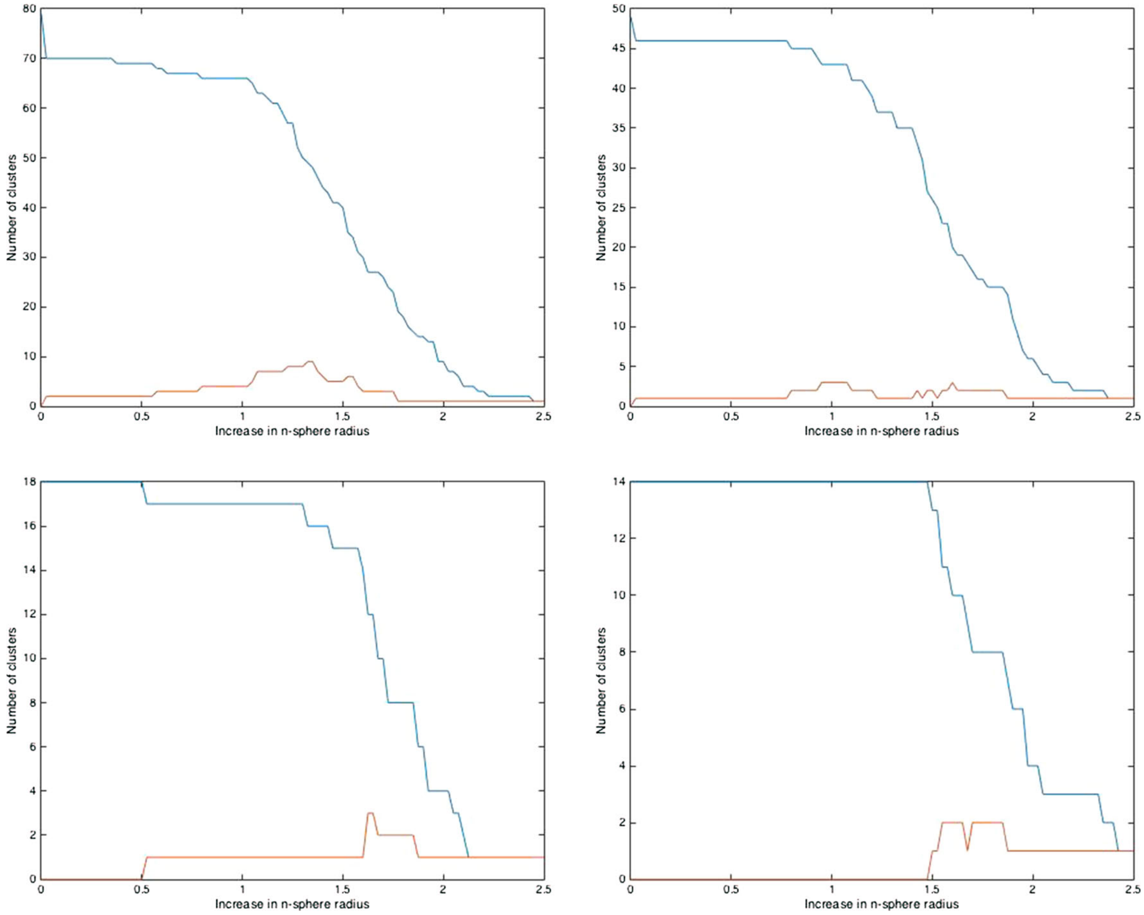

**Fig. 3** Cluster structure of syntactic parameters for the Indo-European (top-left), the Niger-Congo (top-right), the Austronesian (bottom-left), and the Afro-Asiatic (bottom-right) language families. The cluster analysis shows that different language families exhibit different cluster structures, but also that it is still possible to identify common interesting regions of the cluster filter parameters where one can carry out a comparative analysis of the respective persistent topologies. This will be discussed more in detail in § 5.1

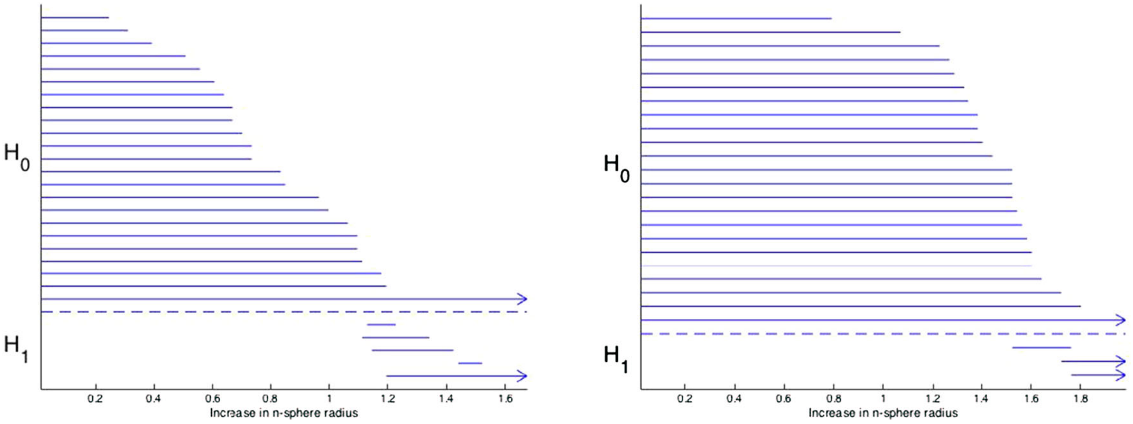

**Fig. 4** Barcode diagram for the Indo-European language family, for cluster filtering value 165 at indices (7, 5, 95) and (10, 5, 95). The persistent topology of the Indo-European language family illustrated here is discussed more in detail in § 5.4. The $(n_1, n_2, n_3)$ indexing set of the diagrams is defined in § 5.2

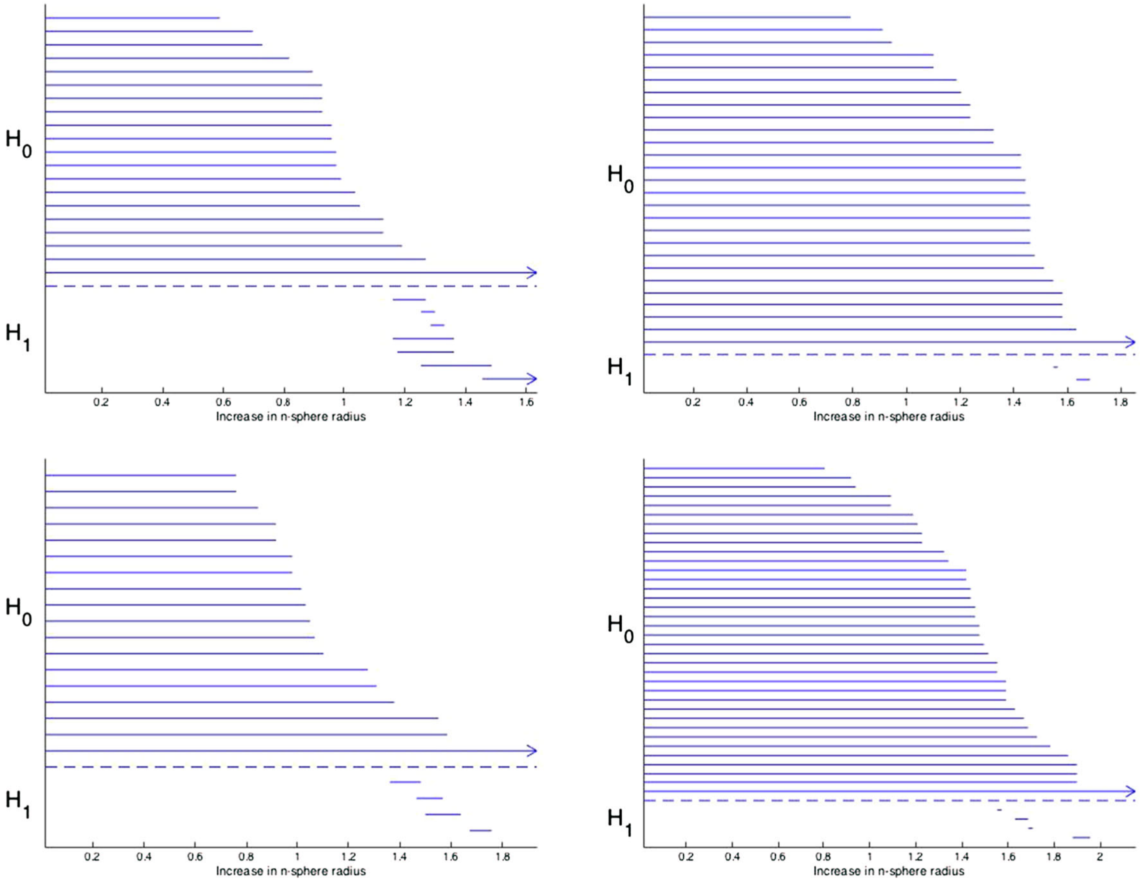

**Fig. 5** Barcode diagram for the Niger-Congo language family, for cluster filtering value 165 and indices (7, 3, 100) and (10, 0, 100), and for cluster filtering value 190 and indices (7, 5, 104) and (10, 0, 104). The persistent topology of the Niger-Congo language family illustrated here is discussed more in detail in § 5.5. The $(n_1, n_2, n_3)$ indexing set of the diagrams is defined in § 5.2

program computing births and deaths of homology generators of the Vietoris–Rips complex. More precisely, the first index (7 or 10) refers to the percent variance divided by 10, while the second index (0, 3 or 5) refers to the percent complete divided by 10. They are discussed above in § 4. The third parameter is the number of steps in Perseus. If present, the additional parameter given by the number after "cluster" is one hundred times the radius used for cluster filtering.

### 5.3 Persistent Topology of the Indo-European Family

We analyze the persistent homology of the syntactic parameters for the Indo-European language family. As shown in Fig. 6, at values (7, 5, 96) and (10, 0, 96) one sees persistent generators of $H_0$ and intervals in the varying $n$-sphere radius, for which nontrivial $H_1$ generators exist. At values (10, 5, 98), as shown in Fig. 6, one sees one persistent generator of $H_1$ and two persistent generators of $H_0$. The existence of a persistent generator for the $H_1$ suggests that there should be a "circle coordinate" description for at part of the syntactic parameters of the Indo-European languages. The fact that there are two persistent generators of $H_0$ in the same diagram indicates two connected components, only one of which is a circle: this component determines which subset of syntactic parameters admits a parameterization by values of a circle coordinate.

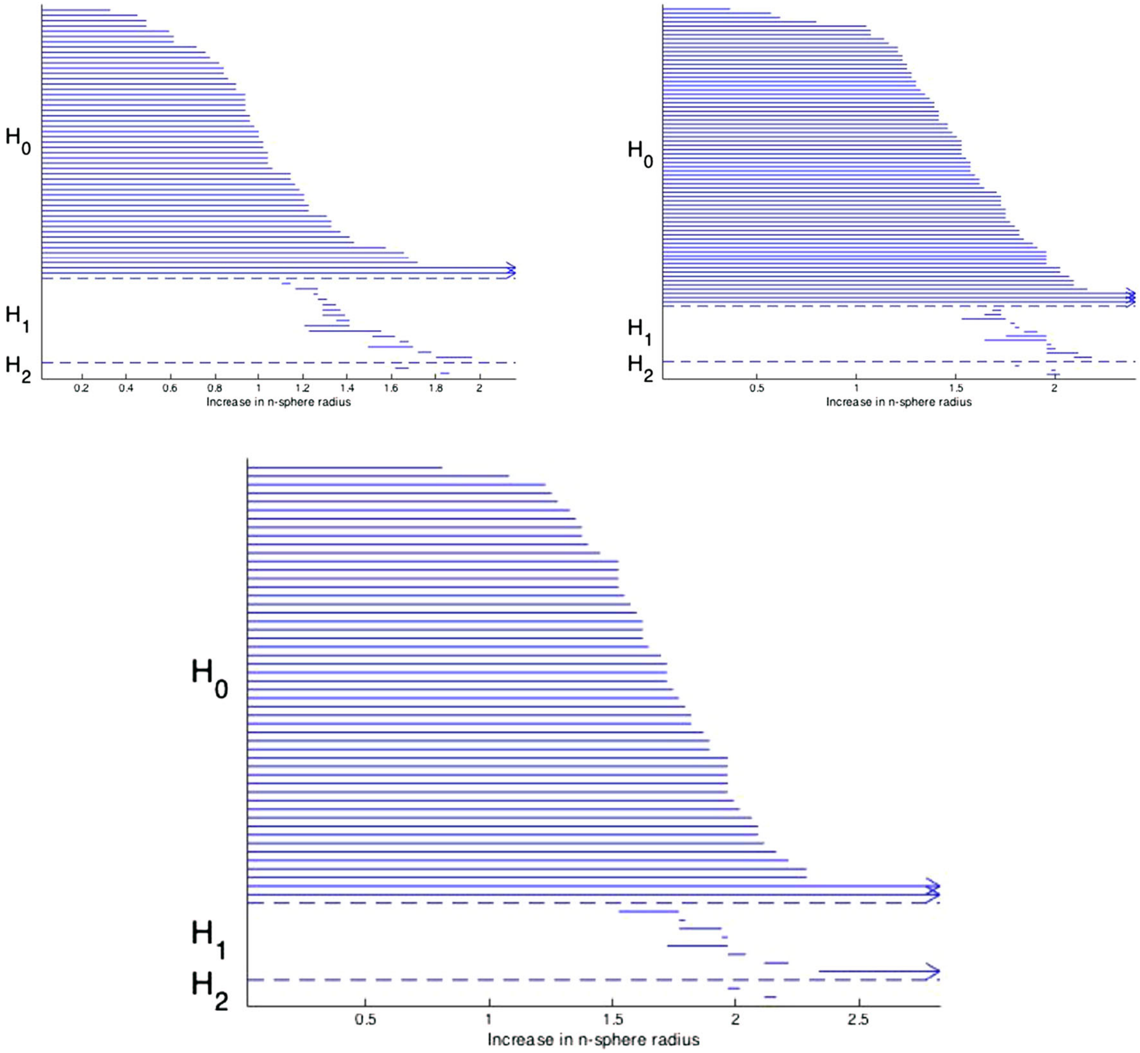

**Fig. 6** Barcode diagram for the Indo-European language family, at indices (7, 5, 96), (10, 0, 96), and (10, 5, 98). The $(n_1, n_2, n_3)$ indexing set of the diagrams is defined in § 5.2

Based on the cluster analysis described in § 5.1 above, we then focus on specific regions of cluster filtering values that are more likely to exhibit interesting topology. For example, for cluster filtering value 165, the results show, respectively, one generator of $H_0$ and one generator of $H_1$, for indices (7, 5, 95), and one generator of $H_0$ and a possibility of two persistent generators of the $H_1$, for indices (10, 5, 95), see Fig. 4. The appearance of persistent generators of the $H_1$ as specific cluster filtering values identifies other groups of syntactic parameters that may admit circle variable parameterizations. What these topological structures in the space of syntactic parameters, and these subsets admitting circle variables description, mean in terms of linguistic theory remains to be fully understood. We analyze some historical-linguistic hypotheses in the following subsection.

### 5.4 Indo-European Persistent Topology and Historical Linguistics

When one analyzes the Indo-European languages at the level of lexical data, it is often argued that the resulting phylogenetic "tree" of the family should not really be a tree, because of the historical influence of French on

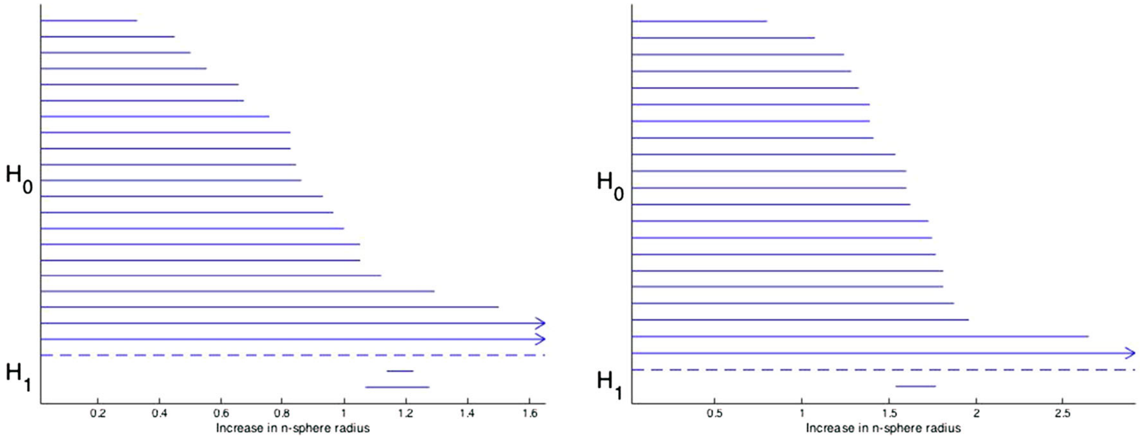

**Fig. 7** Barcode diagram for the Latin+Germanic languages, at indices (7, 5, 129) and (10, 5, 130). The $(n_1, n_2, n_3)$ indexing set of the diagrams is defined in § 5.2

Middle English, which can be viewed as creating a bridge (sometimes referred to as the Anglo-Norman bridge) connecting the Latin (Romance) and Germanic subtrees and introducing non-trivial topology into the Indo-European phylogenetic network. It is well known that the influx of French was extensive at the lexical level, but it is not clear whether one should expect to see a trace of this historical phenomenon when analyzing languages at the level of syntactic structures. In fact, there does not seem to be a clearly detectable syntactic remnant of the Anglo-Norman bridge. It is, however, a natural question to ask whether the non-trivial loop one sees in the persistent topology of syntactic parameters of the Indo-European family may perhaps be related the Anglo-Norman bridge.

A further analysis of the SSWL dataset of syntactic parameters appears to exclude this possibility, confirming what is known in Historical Linguistics, that the Anglo-Normand bridge is only visible at the lexical level. Indeed, we compute the persistent homology using only the Indo-European languages in the Latin and Germanic subtrees. If the persistent generator of $H_1$ were due to the Anglo-Norman bridge one would still find this non-trivial generator when using only this group of languages, while what we find is that the set of Latin and Germanic languages alone carries no non-trivial persistent first homology, see Fig. 7.

In order to understand the nature of the two persistent generators of $H_0$, we separate out the Indo-Iranian subfamily of the Indo-European family, to test whether the two persistent connected components would be related to the natural subdivision into the two main branches of the Indo-European family. Even though the Indo-Iranian branch is the largest subfamily of Indo-European languages, it is much less extensively mapped in SSWL than the rest of the Indo-European family, with only 9 languages recorded in the database. Thus, a topological data analysis performed directly on the Indo-Iranian subfamily is less reliable, but one can gain sufficient information by analyzing the remaining set of Indo-European languages, after removing the Indo-Iranian subfamily. The result is illustrated in Fig. 8. We see that indeed the number of persistent connected components is now just one. This supports the proposal of relating persistent generators of $H_0$ to major subdivisions into historical linguistic subfamilies. Moreover, the persistent generator of the $H_1$ is still present, which shows that the non-trivial first homology is not located in the Indo-Iranian subfamily.

In order to understand more precisely where the non-trivial persistent first homology is located in the Indo-European family, we perform the analysis again after removing the Indo-Iranian languages and also removing the Hellenic branch (including both Ancient and Modern Greek). The resulting persistent topology is illustrated in Fig. 9. By comparing Figs. 8 and 9 one sees that the position of Ancient Greek and the Hellenic branch of the Indo-European family has a direct role in determining the persistent topology. When this subfamily is removed, the number of persistent connected components (generators of $H_0$) jumps from one to three, while the non-trivial single generator of $H_1$ disappears. Although this observation by itself does not provide an explanation of the persistent topology in terms of historical linguistics of the Indo-European family, it points to the fact that, if historical linguistic

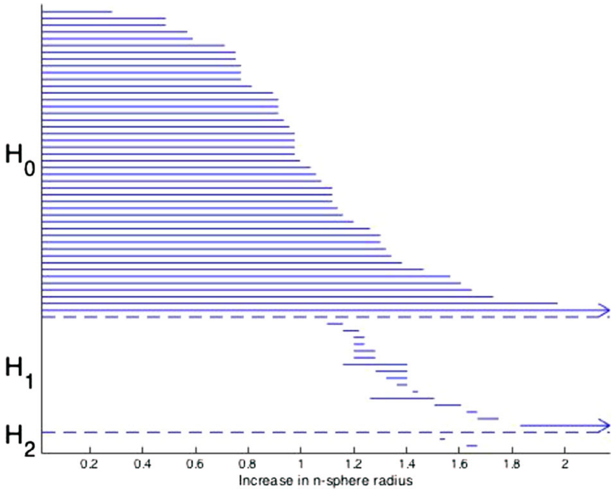

**Fig. 8** Barcode diagram for the Indo-European family with the Indo-Iranian subfamily removed, at indices (7, 5, 97). The $(n_1, n_2, n_3)$ indexing set of the diagrams is defined in § 5.2

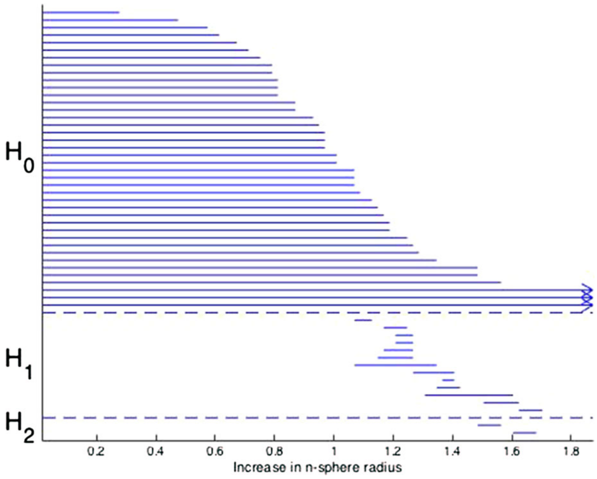

**Fig. 9** Barcode diagram for the Indo-European family with the Indo-Iranian and the Hellenic subfamilies removed, at indices (7, 5, 96). The $(n_1, n_2, n_3)$ indexing set of the diagrams is defined in § 5.2

phenomena are involved in determining the topology, they appear to be related to the role played by Ancient Greek in particular and the Hellenic branch more generally in the historical development of the Indo-European languages. In particular, the circle appears to detect syntactic proximities between the Hellenic branch and other branches of the Indo-European family (including some Slavic languages) that are not seen at the level of the phylogenetic trees.

When performing a more detailed cluster analysis on the Indo-European family, one finds sub-structures in the persistent topology. For instance, as shown in Fig. 4, one sees a possible second generator of the persistent $H_1$ for cluster filtering value 165, with indices (10, 5, 95). These substructures may also be possible traces of other historical linguistic phenomena.

### 5.5 Persistent Topology of the Niger–Congo Family

We perform the same type of analysis on the syntactic parameters of the Niger–Congo language family. The interesting result we observe is that the behavior of persistent homology seems to be quite different for different language families. Figure 10 shows the barcode diagrams for persistent homology at index values (7, 5, 107), (10, 0, 100), and (10, 5, 105), which can be compared with the diagrams of Fig. 6 for the Indo-European family. In the Niger–Congo family, we now see persistent $H_0$ homology, respectively, of ranks 1, 3, and 1 (compare with ranks

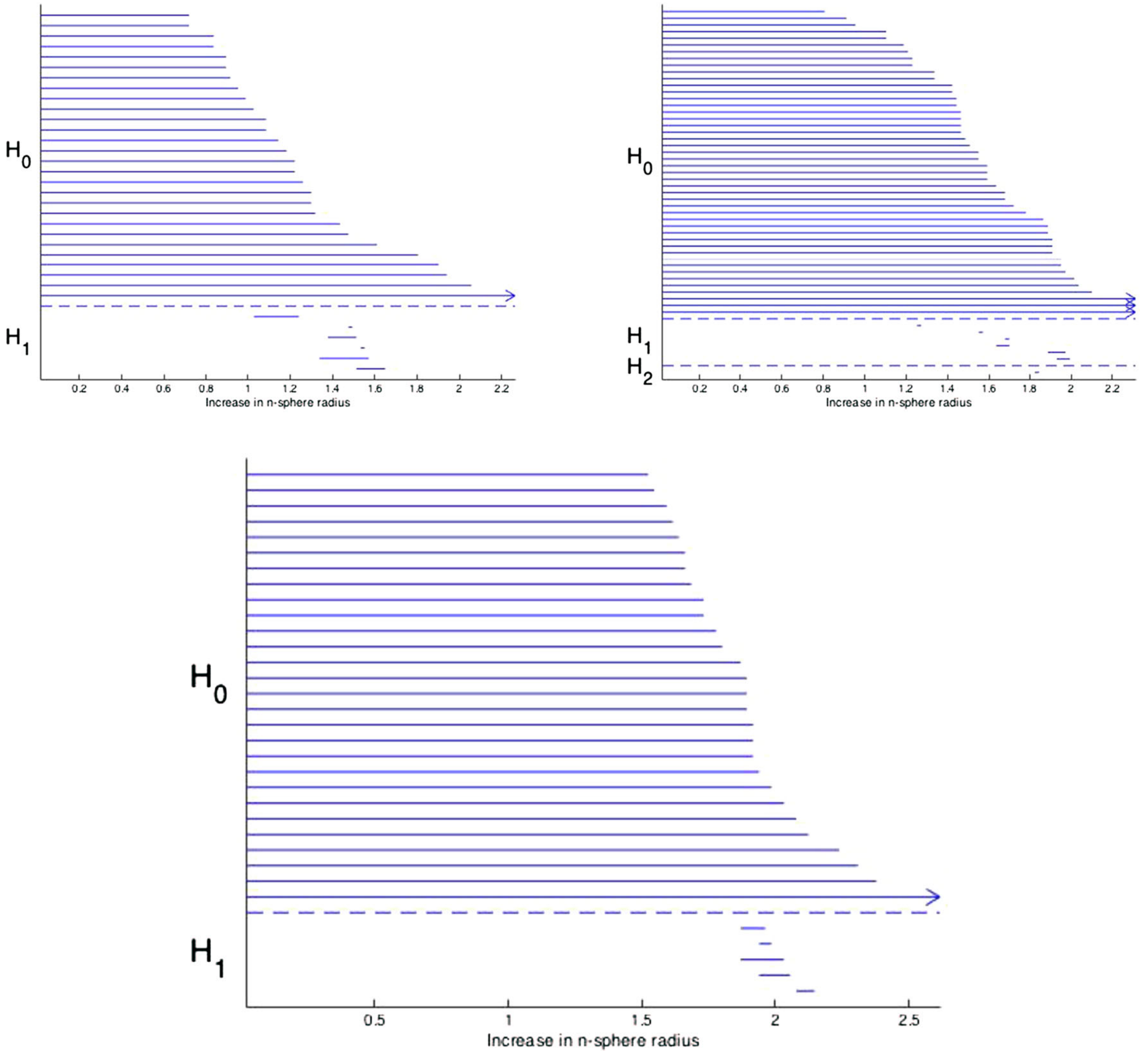

**Fig. 10** Barcode diagram for the Niger-Congo language family, at indices (7, 5, 107), (10, 0, 100), and (10, 5, 105). The persistent topology of the Niger-Congo language family illustrated here is discussed more in detail in § 5.5. The $(n_1, n_2, n_3)$ indexing set of the diagrams is defined in § 5.2.

2, 3, 2 in the Indo-European case). A lower rank in the $H_0$ means fewer connected components in the Vietoris–Rips complex, which seems to indicate that the syntactic parameters are more concentrated and homogeneously distributed across the linguistic family, and less "spread out" into different sub-clusters. The appearance of a range of scales with three persistent components seems to be related to the subdivision into three subfamilies, Mande, Atlantic–Congo, and Kordofanian.

Following the cluster analysis of § 5.1, we also consider the persistent homology for the Niger–Congo family at specific cluster filtering values. While for cluster filtering value 165 and indices (7, 3, 100) one sees one persistent generator of $H_0$ and a possibility of a persistent generator in the $H_1$, cluster filtering value 165 with indices (10, 0, 100), as well as cluster filtering value 190 with indices (7, 5, 104) and (10, 0, 104) only show one persistent generator in the $H_0$.

This persistent homology viewpoint seems to suggest that syntactic parameters within the Niger–Congo language family may be spread out more evenly across the family than they are in the Indo-European case, with a single persistent connected component, whereas the Indo-European ones have two different persistent connected component, one of which has circle topology.

## 6 Further Questions

We showed that methods from topological data analysis, in particular persistent homology, can be used to analyze how syntactic parameters are distributed over different language families. The persistent $H_0$ provides a clustering structure at different scales, which we argue is related to the subdivision into historical subfamilies, while the presence of a persistent $H_1$, which only occurs in certain language families, such as the Indo-European, detects a more subtle structure related to influences at the syntactic level between different branches of the family phylogenetic tree and the role of some specific nodes in the tree in generating these syntactic influences.

We list here some questions that naturally arise from this perspective, which we believe are worthy of further investigation.

(1) To what extent do persistent generators of the $H_0$ (that is, the persistent connected components) of the data space of syntactic parameter correspond to different (sub)families of languages in the historical linguistic sense? The two persistent generators of the Indo-European family appear to match the subdivision into Indo-Iranic and European subfamilies, and (with more uncertainty) the three $H_0$ generators visible at scale $(10, 0, 100)$ in the Congo–Niger family appear to correspond to the subdivision into the Mande, Atlantic–Congo, and Kordofanian subfamilies.

(2) If there is a reliable correspondence between persistent components and historical subfamilies, one can conceivably use the analysis of persistent generators of $H_0$ at different scales as a new method to generate phylogenetic trees of language families. How well do such reconstructions match other phylogenetic reconstructions, based on syntactic data (such as [9, 14]) or on other linguistic data (such as [29])?

(3) Is there a linguistic meaning for the Principal Components of our data set? These determine certain weighted combinations of the syntactic binary variables: how should the weights be regarded from the linguistic viewpoint? Do they have a natural interpretation?

(4) What is the meaning, in historical linguistic terms, of the circle components (persistent generators of $H_1$) in the data space of syntactic parameters of language families? Is it detecting homoplasy in syntax, in a way similar to other forms of homoplasy in Linguistics studied in [24]? Is it detecting syntactic influences across linguistic subfamilies?

(5) There are substructures in the persistent homology that appear at particular scales: how should one interpret those? Is there a historical-linguistic interpretation for the second $H_1$ generator one sees at cluster filtering value 165 and scale $(10, 5, 95)$ in the Indo-European family? Or for the $H_1$ generator one sees with the same cluster filtering, at scale $(7, 3, 100)$ in the Niger–Congo case?

(6) To what extent does persistent topology describe different distribution of syntactic parameters across languages for different linguistic families? The distribuition of syntactic parameters within a given group of languages was studied in [12] and [15] in terms of coding theory: can this approach be related to our approach with persistent homology, perhaps though the methods used in the theory of neural codes, [30], [31]?

**Acknowledgements** This work was performed within the activities of the last author's Mathematical and Computational Linguistics lab and CS101/Ma191 class at Caltech. The last author was partially supported by NSF Grants DMS-1007207, DMS-1201512, DMS-1707882, and PHY-1205440.

## References

1. SSWL Database of Syntactic Parameters: http://sswl.railsplayground.net/
2. Perseus Software Package for Persistent Homology: http://www.sas.upenn.edu/~vnanda/perseus/
3. Chomsky, N.: Lectures on Government and Binding. Foris Publications, Dordrecht (1982)

4. Chomsky, N., Lasnik, H.: The theory of principles and parameters. In: Syntax: An International Handbook of Contemporary Research, pp. 506–569. de Gruyter (1993)
5. Baker, M.: The Atoms of Language. Basic Books, New York (2001)
6. Rizzi, L.: On the format and locus of parameters: the role of morphosyntactic features, preprint (2016)
7. Shopen, T.: Language Typology and Syntactic Description: Volume 1, Clause Structure; Volume 2, Complex Constructions; Volume 3: Grammatical Categories and Lexicon. Cambridge University Press, Cambridge (2007)
8. Galves, C. (ed.): Parameter Theory and Linguistic Change. Oxford University Press, Oxford (2012)
9. Longobardi, G., Guardiano, C.: Evidence for syntax as a signal of historical relatedness. Lingua **119**, 1679–1706 (2009)
10. Haspelmath, M.: Parametric versus functional explanations of syntactic universals. In: The Limits of Syntactic Variation, pp. 75–107. John Benjamins (2008)
11. Haspelmath, M., Dryer, M.S., Gil, D., Comrie, B.: The World Atlas of Language Structures. Oxford University Press, Oxford (2005)
12. Marcolli, M.: Syntactic parameters and a coding theory perspective on entropy and complexity of language families. Entropy **18**(4), 110 (2016)
13. Park, J.J., Boettcher, R., Zhao, A., Mun, A., Yuh, K., Kumar, V., Marcolli, M.: Prevalence and recoverability of syntactic parameters in sparse distributed memories. In: Geometric Science of Information. Third International Conference GSI 2017, vol. 10589, pp. 265–272, Lecture Notes in Computer Science, Springer (2017)
14. Shu, K., Aziz, S., Huynh, V.L., Warrick, D., Marcolli, M.: Syntactic phylogenetic trees. In: Kouneiher, J. (ed.) Foundations of Mathematics and Physics one Century after Hilbert, Springer Verlag. arXiv:1607.02791, to appear
15. Shu, K., Marcolli, M.: Syntactic structures and code parameters. Math. Comput. Sci. **11**(1), 79–90 (2017)
16. Siva, K., Tao, J., Marcolli, M.: Spin Glass Models of Syntax and Language Evolution. arXiv:1508.00504, to appear in Linguistic Analysis
17. Bendor-Samuel, J.: The Niger–Congo Languages: A Classification and Description of Africa's Largest Language Family. University Press of America, Lanham (1989)
18. Manfredi, V., Reynolds, K. (eds.): Niger–Congo Syntax and Semantics. Boston University, African Studies Center, Boston (1995)
19. Shu, K., Ortegaray, A., Berwick, R., Marcolli, M.: Phylogenetics of Indo-European Language Families via an Algebro-Geometric Analysis of their Syntactic Structures, arXiv:1712.01719
20. Carlsson, G.: Topology and data. Bull. Am. Math. Soc. **46**(2), 255–308 (2009)
21. Edelsbrunner, H., Harer, J.L.: Computational Topology: An Introduction. American Mathematical Society, Providence (2010)
22. Ghrist, R.: Elementary Applied Topology. CreateSpace, Seattle (2014)
23. Carlsson, G., Ishkhanov, T., de Silva, V., Zomorodian, A.: On the local behavior of spaces of natural images. Int. J. Comput. Vis. **76**, 1–12 (2008)
24. Warnow, T., Evans, S.N., Ringe, D., Nakhleh, L.: A stochastic model of language evolution that incorporates homoplasy and borrowing. In: Phylogenetic Methods and the Prehistory of Languages, McDonald Institute Monographs (2006)
25. Zomorodian, A., Carlsson, G.: Computing persistent homology. Discrete Comput. Geom. **33**(2), 249–274 (2005)
26. Horak, D., Maletić, S., Rajković, M.: Persistent homology of complex networks. J. Stat. Mech. **2009**, P03034 (2009)
27. Kahle, M.: Random geometric complexes. Discrete Comput. Geom. **45**(3), 553–573 (2011)
28. Pachter, L., Sturmfels, B.: Algebraic statistics for computational biology. Cambridge University Press, Cambridge (2005)
29. Ringe, D., Warnow, T., Taylor, A.: Indo-European and computational cladistics. Trans. Philol. Soc. **100**, 59–129 (2002)
30. Manin, Y.I.: Neural codes and homotopy types: mathematical models of place field recognition. Mosc. Math. J. **15**(4), 741–748 (2015)
31. Curto, C., Itskov, V., Veliz-Cuba, A., Youngs, N.: The neural ring: an algebraic tool for analysing the intrinsic structure of neural codes. Bull. Math. Biol. **75**(9), 1571–1611 (2013)