Math.Comput.Sci. (2021) 15:803–857
https://doi.org/10.1007/s11786-021-00507-2

Mathematics in Computer Science

Phylogenetics of Indo-European Language Families via an Algebro-Geometric Analysis of Their Syntactic Structures

Kevin Shu · Andrew Ortegaray · Robert C. Berwick · Matilde Marcolli

Received: 30 April 2018 / Revised: 20 January 2019 / Accepted: 5 March 2021 / Published online: 15 April 2021

Abstract Using Phylogenetic Algebraic Geometry, we analyze computationally the phylogenetic tree of subfamilies of the Indo-European language family, using data of syntactic structures. The two main sources of syntactic data are the SSWL database and Longobardi's recent data of syntactic parameters. We compute phylogenetic invariants and estimates of the Euclidean distance functions for two sets of Germanic languages, a set of Romance languages, a set of Slavic languages and a set of early Indo-European languages, and we compare the results with what is known through historical linguistics.

Keywords Phylogenetic algebraic geometry · Syntactic parameters · Historical linguistics · Phylogenetic trees · Indo-European languages

Mathematics Subject Classification 91F20 · 92D15 · 05C85 · 60J10

1 Introduction

The use of commutative algebra and algebraic geometry in the study of phylogenetic trees and networks was developed in recent years in the context of biological applications, see [35,36]. We argue in this paper that these

K. Shu . A. Ortegaray . M. Marcolli (✉)
California Institute of Technology, Pasadena, USA
e-mail: matilde@caltech.edu;
matilde@math.utoronto.ca

K. Shu
e-mail: kshu@caltech.edu

A. Ortegaray
e-mail: aortegar@caltech.edu

R. C. Berwick
Massachusetts Institute of Technology, Cambridge, USA
e-mail: berwick@csail.mit.edu

M. Marcolli
Perimeter Institute for Theoretical Physics, Waterloo, Canada

M. Marcolli
University of Toronto, Toronto, Canada

methods have advantages over the other methods of phylogenetic reconstruction, such as Hamming distance and neighbor-joining, when applied to the computational study of phylogenetic trees of world languages based on syntactic data. Computational studies of phylogenetics in linguistics have been carried out recently in [4,52], using lexical and morphological data and in [27,28] using syntactic data.

The main advantages of the algebro-geometric approach presented here can be summarized as follows.

(1) The use of Phylogenetic Algebraic Geometry to select a best candidate tree avoids some of the well known possible problems (see Chapter 5 of [51]) that can occur in phylogenetic reconstructions based on Hamming distance and neighbor-joining methods. While such methods were used successfully in phylogenetic inference using syntactic data in [27] and [28], we argue that the geometric methods provide additional useful information, as explained below.

(2) Phylogenetic Algebraic Geometry associates an actual *geometric object* to a best candidate phylogenetic tree T, together with a boundary probability distribution at the leaves $P = (p_{i_1 \dots i_n})$ derived from the data. This geometric object consists of a pair $(V_T, x_{T,P})$ of an algebraic variety V_T, which depends on the tree topology, and a point $x_{T,P} \in V_T$ on it, which depends on both the tree T and the boundary distribution P. Unlike what happens with other phylogenetic methods that only provide a best candidate tree T, the geometry $(V_T, x_{T,P})$ contains more information: the position of the point P on the variety V_T encodes information about the distribution of the binary syntactic features across the language family. For example, one can have different language families with topologically equivalent phylogenetic trees. In this case one obtains two different points on the same variety V_T whose relative positions encode in a quantitative geometric way the difference between how the evolution of syntactic feature happened historically in the two families.

(3) The point $x_{T,P}$ is constrained to lie on the locus of real points $V_T(\mathbb{R})$ of the complex algebraic variety V_T, and in particular on the sublocus $V_T(\mathbb{R}_+)$ of nonnegative real coordinates, since it is defined by a probability distribution. In several cases, especially when analyzing sufficiently small trees, V_T turns out to be a classical and well studied algebraic variety, as in the case of the Secant varieties of Segre embeddings of products of projective spaces that we encounter in this paper. In such cases, there are usually well understood and interesting geometric subvarieties of V_T and one can gain further insight by understanding when the point $x_{T,P}$ lies on some of these subvarieties, in addition to being contained in the real locus. For example, this may suggest compatibility of the boundary distribution P with respect to certain splitting of the tree into subfamilies and subtrees, which may provide additional information on the underlying historical linguistics.

(4) The algebro-geometric method is compatible with admixtures and with phylogenetic networks that are not necessarily trees. The algebraic varieties involved in this setting are different from the phylogenetic varieties of trees V_T discussed here, but they are analyzed with a similar method. Results on topological analysis of data of syntactic structures (see [41]) indicate the presence of nontrivial cycles (first homology generators) in certain language families. This can be seen as supporting evidence for the use of networks that are not trees for phylogenetic analysis. The algebro-geometric formalism necessary to the discussion of more general phylogenetic networks is discussed in [37] and [9].

1.1 Binary Variables and Syntactic Structure

The idea that the possible syntactic structure of human languages is governed by certain basic binary variables, or syntactic parameters, is one of the fundamental ideas underlying the Principles and Parameters model in linguistics, originally introduced by Chomsky [10,12]. The notion of syntactic parameter underwent successive theoretical reformulation in the context of more recent minimalist models [11], but the main underlying conceptual idea remains unchanged. A recent detailed overview of the state of ongoing research in comparative generative grammar on the topic of syntactic parameters can be found in the collection of papers in the volume [22]. An introduction to syntactic parameters aimed at a general audience with no prior linguistics background is given in [3].

Interesting questions regarding syntactic parameters include identifying a minimal set of independent variables completely determining a language's syntax and obtaining an explicit and complete description of the dependencies

that exist among the known parameters. A rough analogy is that the set of syntactic parameters forms a kind of "basis set" spanning the space of possible human languages (alternatively, grammars, since we are attempting to describe language structure). Each choice of values for the parameters in this basis set fixes a distinct possible (presumably learnable) human language. Typically, it is assumed that the parameter values can be learned from data available from positive example sentences presented to a language learner (i.e., a child). These binary variables describing syntactic structures can roughly be thought of as yes/no answers to questions about whether certain constructions are possible in a given language or not. For a more precise description of parameters as instructions for triggering syntactic operations see [44].

From a more precise mathematical perspective one can view the question of identifying dependencies between syntactic parameters as trying to identify the correct "manifold of syntax" inside a large ambient space of binary variables, in the same sense as constraints on a physical system determine the manifold structure of its configuration space. Any existing relation between syntactic parameters determines a locus inside the space of all possible binary values of these syntactic variables where the syntactic data of the actual human languages are constrained to lie. Since identifying relations between syntactic parameters is an open problem, the resulting geometry cut out by these relations is presently unknown. While the problem of the "geometry of syntax" in itself is not the main focus of the present paper, the issue of dependencies between syntactic variables is relevant, because the phylogenetic models we will be discussing are typically based on assuming that variables evolving according to a Markov process on a tree behave like independent identically distributed (i.i.d.) random variables. While this assumption is good enough to draw some reasonable linguistic conclusions, in a more refined analysis one would like to identify the extent to which relations between syntactic parameters may cause deviations from this hypothesis. This problem will be discussed elsewhere [17].

There are two existing databases of syntactic structures of world languages that we use in this paper: the SSWL database [49] and the data of syntactic parameters collected by Giuseppe Longobardi and the LanGeLin collaboration. The binary variables recorded in the SSWL database should not be regarded, from the linguistics perspective as genuine syntactic parameters, although they still provide a very useful collection of binary variables describing different features of syntactic structures of world languages. The variables recorded in the SSWL database include a set of 22 binary variables describing word order properties, *01*–Subject Verb,. . ., *22*–Noun Pronomial Possessor, a set of 4 binary variables *A01–A04* describing relations of adjectives to nouns and degree words, a variable *AuxSel01* about the selection of auxiliary verbs, variables *C01–C04* still related to word order properties on complementarizer and clause and adverbial subordinator and clause, *N201–N211* variables on properties of numerals, *Neg01–Neg14* variables on negation, *OrderN301–OrderN312* on word order properties involving demostratives, adjectives, nouns, and numerals, *Q01–Q15* regarding the structure of questions, *Q16Nega–Q18Nega* and *Q19NegQ–Q22NegQ* on answers to negative questions, *V201-V202* on declarative and interrogative Verb-Second, *w01a–w01c* indefinite mass nouns in object position, *w02a–w02c* definite mass nouns in object position, *w03a–w03d* indefinite singular count nouns in object position, *w04a–w04c* definite singular count nouns in object position, *w05a–w05c* indefinite plural count nouns in object position, *w06a–w06c* definite plural count nouns in object position, *w07a–w07d* nouns with (intrinsically) unique referents in object position, *w08a–w08d* proper names in object position, *w09a–w09b* order of article and proper names in object position, *w10a–w10c* proper names modified by an adjective in object position, *w11a–w11b* order of proper names and adjectives in object position, *w12a–w12f* order of definite articles and nouns in object position, *w20a–w20e* singular count nouns in vocative phrases, *w21a–w21e* proper nouns in vocative phrases, *w22a–w22e* plural nouns in vocative phrases. A detailed description of each of these binary variables can be found on the online site of the SSWL database [49]. While these are certainly not considered to be an exhaustive list of binary variables associated to syntax, they contain a considerable amount of information on the variability of syntactic structures across languages.

The LanGeLin data of Longobardi record a different set of syntactic features, which are independent of the SSWL data. These variables should be regarded as genuine syntactic parameters and are based on the general Modularized Global Parameterization approach developed by Longobardi [24,26], that considers reasonably large sets of parameters within a single module of grammar, and their expression across a large number of languages. The LanGeLin data presented in [24] that we use here include 91 parameters affecting the Determiner Phrases structure.

The full list of the LanGeLin syntactic parameters used in this paper is reported in Appendix D, reproduced from Appendix A of [21].

Unlike the SSWL data, which do not record any explicit relations between the variables, many explicit relations between the Longobardi syntactic parameters are recorded in the LanGeLin data. A more detailed analysis of the relations in the LanGeLin data is given in [21] and in [34]. In our analysis here we have removed those parameters in the LanGeLin data that are explicitly dependent upon the configuration of other parameters.

1.2 Related Work

A long-standing, familiar approach to linguistic phylogenetics is grounded on the use of lexical (including phonemic) features; see, e.g., [52] for a survey of phylogenetic methods applying such features on a carefully analyzed Indo-European dataset. More recently, other researchers have suggested alternatives to bypass issues with lexical items, such as the non-treelike behavior of lexical diffusion, sometimes rapid and different time scales for lexical change, and the like. For example, Murawaki [32] used linguistic typological dependencies such as word order (OV vs. VO, in the Greenbergian sense) or grammar type (synthetic vs. analytic), in order to build phylogenies over longer time scales and across widely different languages. Murawaki's approach computes latent components from linguistic typological features in the *World Atlas of Languages*, (WALS) and then feeds these into phylogenetic analysis. Longobardi and colleagues have pursued a detailed linguistically-based analysis of, e.g., Noun Phrases (so-called Determiner structure) across many different Western European languages to develop a fine-grained explicit parametric analysis of what distinguishes each of these languages from the others, see [27] and subsequent work including the more recent [29]. In effect, this is a "hand-tooled" version of a statistical, principal-components like approach. They have used Jacquard distance metrics as the measure to feed into conventional distance-based phylogenetic programs. The approach presented in the current work differs from either of these and from other more familiar phylogenetic methods applied to linguistic datasets (such as maximum likelihood or Bayesian approaches) in that it adopts a different approach to the structure of the phylogenetic space itself, rather than relying on conventional methods, while retaining the non-lexical, typological information as the basis for describing the differences among languages.

1.3 Comments on the Data Sets

The two databases used in our analysis, namely the SSWL database [49] and the recent set of data published by Longobardi and collaborators [24], are currently the only existing extensive databases of syntactic structures of world languages. Therefore any computational analysis of syntax necessarily has to consider these data.

In the process of evaluating phylogenetic trees via the algebro-geometric method, we also perform a comparative analysis of the two databases of syntactic variables that we use. As the extended version of the Longobardi dataset has only recently become available [24], a comparative analysis of this dataset has not been previously considered, so the one reported here is novel. Other methods of comparative analysis of these two databases of syntactic structures will be discussed elsewhere. In the cases analyzed here we see specific examples (such as the second set of Germanic languages we discuss) where Longobardi's database appears to be more reliable for phylogenetic reconstructions than the SSWL data, even though the latter dataset is larger.

1.4 Phylogenetics and Syntactic Data

The use of syntactic data for phylogenetic reconstruction of language families was developed in previous work of Longobardi and collaborators, [27,28], see also [25,26]. Computational phylogenetic reconstructions of language family trees based on lexical and morphological data were also obtained in [4,33,52]. It is well known that the use of lexical data, in the form of Swadesh lists, is subject to issues related to synonyms, loan words, and false positives, that may affect the measure of proximity between languages. Morphological information is much more robust, but its encoding into binary data is not always straightforward. Syntactic data, on the other hand, are usually classified in terms of binary variables (syntactic parameters), and provide a robust information about language structure.

Thus, we believe that syntactic data should be especially suitable for the use of computational methods in historical linguistics.

In [47] it was shown that, when using syntactic data of the SSWL database [49] with Hamming distances and neighbor-joining methods to construct linguistic phylogenetic trees, several kinds of errors typically occur. These are mostly due to a combination of two main factors:

- the fact that at present the SSWL data are very non-uniformly mapped across languages;
- errors propagated by the use of neighbor-joining algorithms based on the Hamming distance between the strings of syntactic variables recorded in the SSWL data.

An additional source of problems is linguistic in nature, namely the existence of languages lying in historically unrelated families that can have greater similarity than expected at the level of their syntactic structures. Another possible source of problems is due to the structure of the SSWL database itself, where the syntactic binary variable recorded are not what linguists would consider to be actual syntactic parameter in the sense of the Principles and Parameters model [10,12], see also [44]: there are conflations of deep and surface structures that make certain subsets of the syntactic variables of the SSWL data potentially problematic from the linguistic perspective. However, it was also shown in [47] that several of these problems that occur in a naive use of computational phylogenetic methods can be avoided by a more careful analysis. Namely, some preliminary evidence is given in [47] that, when a naive phylogenetic reconstruction applied simultaneously to the entire SSWL database is replaced by a more careful analysis applied to smaller groups of languages that are more uniformly mapped in the database, the phylogenetic invariants of Phylogenetic Algebraic Geometry can identify the correct phylogenetic tree, despite the imperfect nature of the SSWL data. The method of Phylogenetic Algebraic Geometry that we refer to here was developed in [35,36] for applications to mathematical biology, see also a short survey in [5].

In the present paper we focus on certain subfamilies of the Indo-European language family, in particular the Germanic languages, the Romance languages, and the Slavic languages. We apply the Phylogenetic Algebraic Geometry method, by computing the phylogenetic invariants for candidate trees, and the Euclidean distance function. We compare the results obtained by applying this method to the SSWL data and to a more recent set of data of syntactic parameters collected by Longobardi [24], which are a largely extended version of the data previously available in [27].

We list here the specific historical linguistics settings that we analyze in this paper.

1.5 The Germanic Family Tree

We consider the following two sets of Germanic languages:

(1) $\mathcal{S}_1(G) = \{\text{Dutch, German, English, Faroese, Icelandic, Swedish}\}$
(2) $\mathcal{S}_2(G) = \{\text{Norwegian, Danish, Icelandic, German, English, Gothic, Old English}\}$.

The first one only consists of modern languages, while in the second one we have included the data of the two ancient languages Gothic and Old English. We analyze the first set $\mathcal{S}_1(G)$ with the SSWL data, and we analyze the second set first using the new Longobardi data and then using the SSWL data. In both cases we first generate candidate trees using the software package PHYLIP [40], then using the Phylogenetic Algebraic Geometry method we compute the phylogenetic invariants and an estimate of the Euclidean distance function for these candidate trees and we select the best candidate.

For sufficiently small trees one can expect that other methods, including more conventional Bayesian analysis, would be able to identify the correct candidate tree. However, we see here in specific examples that the algebro-geometric method performs at least better than standard phylogenetic packages like PHYLIP when applied to the same data.

Given the large number of alternative phylogenetic methods, why use PHYLIP as a baseline? There are two main reasons. First of all, PHYLIP is selected here as an example of a well known and widely used phylogenetic

package, hence it is an easy baseline for comparison. Moreover, we use PHYLIP to preselect a set of candidate trees because likewise parsimony method is a standard starting point for Bayesian analysis, although maximum likelihood inference is generally regarded as a more reliable method.

The estimates we consider here are based on the evaluation of phylogenetic invariants and on estimates of Euclidean distance. A maximum likelihood degree, which counts the critical points of the likelihood function on determinantal varieties, can in principle also be computed, see [23], but only in sufficiently small cases. Although there are cases (such as Gaussian models) where the maximum likelihood degree and the Euclidean distance degree match, there are also many examples where these solutions are different, as shown in [13].

We show that, for the set $\mathcal{S}_1(G)$, the phylogenetic invariants suggest the correct tree among the six candidates generated by PHYLIP, which is confirmed via the estimate of the Euclidean distance. The topology of this tree correctly corresponds to the known historical subdivision of the Germanic languages into West Germanic and North Germanic and the relative proximity of the given languages within these subtrees. In this sense the algebro-geometric method applied to a baseline dataset can be confirmed, always a key step in advancing a novel phylogenetic approach as [52] note.

For the other set $\mathcal{S}_2(G)$ of seven languages, which are common to both databases, we also find that the phylogenetic invariants computed on a subset of the Longobardi syntactic data point to the correct best candidate tree, which is confirmed by a lower bound estimate of the Euclidean distance. With the SSWL data the phylogenetic invariants computed with respect to the ℓ^1 norm still identify the historically correct tree as the best candidate, but not when computed with respect to the ℓ^∞ norm. This confirms in our setting a general observation of [8] on the better reliability of the ℓ^1 norm in the computation of phylogenetic invariants. We see here an example where the lower bound on the Euclidean distance correctly excludes some of the candidates, but fails to assign the smallest lower bound to the best tree. This different behavior of the Longobardi and the SSWL data on this set of languages presumably reflects the presence of a large number of dependencies in the SSWL variables.

In the last section of the paper we discuss a possible issue of the direct application of this algebraic phylogenetic method to syntax, which is caused by neglecting relations between syntactic parameters and treating them, in this model, like independent random variables. We suggest possible ways to correct for these discrepancies, which will be analyzed in future work. We expect that such discrepancies may be resolved by a better approach taking syntactic relations into account.

1.6 The Romance Family Tree

The case of the Romance languages is an interesting example of the limitations of these methods of phylogenetic reconstructions. We considered as set of languages Latin, Romanian, Italian, French, Spanish, and Portugues, and we used a combination of the SSWL and the Longobardi data, which are independent sets of data. We find that PHYLIP produces a unique candidate tree, which is however not the one that is considered historically correct. We compute the phylogenetic invariants and the Euclidean distance for both the PHYLIP tree and the historically correct tree. The phylogenetic invariants computed with respect to the ℓ^1 norm identify the historically correct tree as the favorite candidate, while they do not give useful information when computed in the ℓ^∞ norm. The estimate of the Euclidean distance also favors the historically correct tree over the PHYLIP candidate tree.

1.7 The Slavic Family Tree

We also analyze with the same method the phylogenetic tree of a group of Slavic languages for which we use a combination of SSWL data and the data of [27]: Russian, Polish, Slovenian, Serb-Croatian, Bulgarian. For this set of languages, PHYLIP applied to the combined syntactic data produces five candidate trees with inequivalent topologies. Using the phylogenetic invariants computed with the ℓ^1 norm we identify the historically correct tree as the best candidate, while the computation in the ℓ^∞ norm does not select a unique best candidate. The lower bound estimate of the Euclidean distance also correctly selects the linguistically accurate tree.

1.8 The Early Indo-European Branchings and the Indo-European Controversy

The use of computational methods in historical linguistics has been the focus of considerable attention, and controversy, in recent years, due to claims made in the papers [6,18] regarding the phylogenetic tree of the Indo-European languages, based on a computational analysis of trees obtained from distances between binary data based on lexical lists and cognate words. While this method of computational analysis of language families has been considered in various contexts (see [16] for a collection of contributions), the result announced in [6,18] appeared to contradict several results obtained by historical linguists by other methods, hence the ensuing controversy, see [39]. For comparison, a different reconstruction of the Indo-European tree, carried out by computational methods that incorporate lexical, phonological, and morphological data, was obtained by Ringe, Warnow, and Taylor [43]. Neither of these computational analysis makes any use of syntactic data about the Indo-European languages.

We focus here on some specific issues that occur in the phylogenetic tree of [6] compared with that of [43]:

- The relative positions of the Greco-Armenian subtrees;
- The position of Albanian in the tree;
- The relative positions of these languages with respect to the Anatolian-Tocharian subtrees.

This means that we neglect several other branches of the Indo-European tree analyzed in [6] and in [43] and we focus on a five-leaf binary tree with leaves corresponding to the languages: Hittite, Tocharian, Albanian, Armenian, and Greek. We will consider the tree topologies for this subset of languages resulting from the trees of [6] and [43] and we will select between them on the basis of Phylogenetic Algebraic Geometry.

The set of languages considered here (Hittite, Tocharian, Albanian, Armenian, Greek) are listed in the SSWL database [49], while not all of them are present in the Longobardi data [24]. Thus, in this case we have to base our analysis on the SSWL data. With the exception of Armenian and Greek, which are extensively mapped in the database, the remaining languages (especially Tocharian and Hittite) are very poorly mapped, and the set of parameters that are completely mapped for all of them is very small, hence the resulting analysis should not be considered very reliable, due to this significant problem.

Nonetheless, we compute the phylogenetic invariants for the Gray-Atkins tree and for the Ringe–Warnow–Taylor tree and we also compute the Euclidean distance function to the relevant phylogenetic algebraic variety. We find that, while the evaluation of the phylogenetic invariants with the ℓ^∞ norm does not give useful information, the evaluation in the ℓ^1 norm favors the linguistically more accurate Ringe–Warnow–Taylor tree. Similarly the estimate of the Euclidean distance selects the same Ringe–Warnow–Taylor tree.

The Gray-Atkins tree is *not* the one generally agreed upon by linguists, while the Ringe–Warnow–Taylor tree is considered linguistically more reliable. A more recent discussion of the early Indo-European tree, which is also considered linguistically very reliable, can be found in [2]. However, the part of the tree of [2] that we focus on here agrees with the one of [52] (though the position of Albanian is not explicitly discussed in [2]), hence we refer to [52] in our analysis.

2 Phylogenetic Algebraic Varieties and Invariants

Before we proceed to the analysis of the two sets of languages listed above, we recall briefly the notation and the results we will be using from Phylogenetic Algebraic Geometry, see [1,35,36]. We also discuss the limits of the applicability of this method to syntactic data of languages and some approaches to improve the method accordingly.

In order to apply the algebro-geometric approach, we think of each binary syntactic variable as a dynamical variable governed by a Markov process on a binary tree. These binary Markov processes on trees generalize the Jukes–Cantor model, in the sense that they do not necessarily assume a uniform distribution at the root of the tree. The model parameters (π, M^e) consist of a probability distribution $(\pi, 1-\pi)$ at the root vertex (the frequency of expression of the 0 and 1 values of the syntactic binary variables at the root) and bistochastic transition matrices

$$M^e = \begin{pmatrix} 1-p_e & p_e \\ p_e & 1-p_e \end{pmatrix}$$

along the edges.

For a binary tree with n leaves, the boundary distribution $P = (p_{i_1,\dots,i_n})$ counts the frequencies of the occurrences of binary vectors $(i_1, \dots, i_n) \in \{0, 1\}^n$ of values of the binary syntactic variables for the languages $\{\ell_1, \dots, \ell_n\}$ at the leaves of the tree. This boundary distribution is the marginal distribution obtained after marginalizing over the internal nodes of the tree. If N is the total number of syntactic binary variables available in the database (counting only those that are completely mapped for all the n languages considered) and $n_{i_1,\dots,i_n}$ is the number of occurrences of the binary vector $(i_1, \dots, i_n)$ in the list of values of the N syntactic variables for these n languages, then the frequencies in P are given by

$$p_{i_1,\dots,i_n} = \frac{n_{i_1,\dots,i_n}}{N}.$$

The boundary distribution is a polynomial function of the model parameters

$$p_{i_1,\dots,i_n} = \Phi(\pi, M^e) = \sum_{w_v \in \{0,1\}} \pi_{w_{v_r}} \prod_e M^e_{w_{s(e)}, w_{t(e)}}, \tag{2.1}$$

with a sum over "histories", that is, paths in the tree. This determines a polynomial map of affine spaces

$$\Phi_T : \mathbb{A}^{4n-5} \to \mathbb{A}^{2^n}, \tag{2.2}$$

where $4n - 5$ is the number of model parameters for a binary tree T with n-leaves and binary variables. Dually, the kernel of the map of polynomial rings

$$\Psi_T : \mathbb{C}[z_{i_1,\dots,i_n}] \to \mathbb{C}[x_1, \dots, x_{4n-5}] \tag{2.3}$$

defines the phylogenetic ideal $\mathcal{I}_T$. This corresponds geometrically to the phylogenetic algebraic variety V_T.

It is proved in [1] that, for these Markov models on trees with binary variables that generalize the Jukes–Cantor model, the phylogenetic ideal $\mathcal{I}_T$ is generated by all the 3×3-minors of all the *flattenings* of the tensor $P = (p_{i_1,\dots,i_n})$. There is one such flattening for each internal edge of the binary tree, where each internal edge corresponds to a subdivision of the leaves into a disjoint union of two sets of cardinality r and $n - r$. The flattening is a $2^r \times 2^{n-r}$ matrix defined by setting

$$\mathrm{Flat}_{e,T}(P)(u, v) = P(u_1, \dots, u_r, v_1, \dots, v_{n-r}), \tag{2.4}$$

where P is the boundary distribution. The terminology corresponds to the fact that an n-tensor P is "flattened" into a collection of 2-tensors (matrices).

These generators of the phylogenetic ideal can then be used as a test for the validity of a candidate phylogenetic tree. If the tree is a valid phylogenetic reconstruction, then the boundary distribution $P = (p_{i_1,\dots,i_n})$ should be a zero of all the polynomials in the phylogenetic ideal (or very close to being a zero, allowing for a small error margin).

In the case of the binary Jukes–Cantor model, where one assumes a uniform root distribution, there are additional invariants, as shown in [50]. For the purpose of linguistic applications it is more natural to work with the general binary Markov models described above, where the root distribution $(\pi, 1 - \pi)$ is not assumed to be uniform, than with the more restrictive Jukes–Cantor model. Indeed, there is no reason to assume that parameters at the root of a language phylogenetic tree would have equal frequency of expression of 0 and 1: the overall data on all languages, ancient and modern, contained in the available database show a clear prevalence of parameters that are expressed (value 1) rather than not. (This point was discussed in some detail in [48].)

2.1 Phylogenetic Invariants

The Allman–Rhodes theorem [1] shows that the generators ϕ_T of the phylogenetic ideal $\mathcal{I}_T$ are given by the minors $\det(M)$ of all the size 3×3-submatrices M of the flattening matrices $\mathrm{Flat}_{e,T}$, with e ranging over the internal edges of T.

In the following, we denote by $\mathcal{M}^{(3)}_{e,T}$ the set of all 3×3 submatrices of the flattening matrix $\mathrm{Flat}_{e,T}$, by $\mathcal{M}^{(3)}_T := \cup_{e \in E(T)} \mathcal{M}^{(3)}_{e,T}$ and by $\mathcal{D}^{(3)}_T := \{\det(M) \mid M \in \mathcal{M}^{(3)}_T\}$. We will also use the notation $\mathcal{M}^{(3)}(A)$ for the set of 3×3 submatrices of a given matrix A, and $\mathcal{D}^{(3)}(A) := \{\det(M) \mid M \in \mathcal{M}^{(3)}(A)\}$.

To every candidate tree, one can also associate a computation of a discrepancy that measures how much the polynomials ϕ_T fail to vanish at the point P. This can be done using different kinds of norms. Generally, one can use either the ℓ^∞ norm and obtain an expression of the form

$$\|\phi_T(P)\|_{\ell^\infty} = \max_{M \in \mathcal{M}^{(3)}_T} |\det(M(P))|,$$

which we write equivalently in the following shorthand notation as

$$\|\phi_T(P)\|_{\ell^\infty} = \max_{\phi \in \mathcal{D}^{(3)}_T} |\phi(P)|,$$

where the expression $|\phi(P)|$ stands for the absolute value of the determinant of the 3×3-minor evaluated at the boundary distribution P. It is also natural to use the ℓ^1 norm and compute

$$\|\phi_T(P)\|_{\ell^1} = \sum_{M \in \mathcal{M}^{(3)}_T} |\det(M(P))|,$$

equivalently written in the rest of the paper as

$$\|\phi_T(P)\|_{\ell^1} = \sum_{\phi \in \mathcal{D}^{(3)}_T} |\phi(P)|.$$

One can expect that the ℓ^∞ norm will be a very weak invariant, because taking the maximum loses a lot of information contained in the phylogenetic invariants $\phi_T(P)$. Indeed, this turns out to be the case. As analyzed in detail in [8], the ℓ^1 norm is a more refined and reliable way to identify best phylogenetic trees on the basis of the computation of phylogenetic invariants than the ℓ^∞ norm. We will see several explicit examples in the following sections where the ℓ^∞ norm does not provide useful information to identify the correct candidate tree, while the ℓ^1 norm of the phylogenetic invariants correctly identifies the unique best candidate tree.

For the best candidate tree T, the values of $\|\phi_T(P)\|_{\ell^\infty}$ and $\|\phi_T(P)\|_{\ell^1}$ will in general be small but still non-zero. It is possible that these non-zero values may partly reflect a small deviation from Markov evolution. Namely, the observed distribution P of the syntactic parameters of the languages at the leaves of the tree may differ from a distribution obtained by the evolution of i.i.d. random variables via a Markov model on the tree.

One of the important points we wish to investigate in the longer term is how relations between syntactic parameters affect their behavior as random variables in dynamical models of language change and evolution. To that purpose, we can regard the values of phylogenetic invariants as a possible numerical indicator of discrepancies from the standard i.i.d. Markov model assumption. As mentioned in the introduction, the presence of dependencies between syntactic parameters is expected to cause at least some small deviations from the dynamics of an actual i.i.d. Markov model. We do not analyze in the present paper how possible models of parameter dependencies affect the dynamics and may be reflected in the value of the phylogenetic invariants. A more careful analysis of the Markov hypothesis will appear elsewhere [17].

2.2 Euclidean Distance

As a way to compare different candidate trees and select the best possible candidate, one can use the Euclidean distance, in an ambient affine space, between the point P given by the boundary distribution and the variety V_T associated to the candidate tree T. The tree realizing the smallest distance will be the favorite candidate.

It is not always possible to compute the Euclidean distance exactly, but it can sometimes be estimated, as we will discuss more explicitly in Sects. 3.6 and 3.11. We will compute Euclidean distances from certain Segre and secant varieties, namely determinantal varieties of rank one and two, for which a direct computation is possible. In some particular cases, like the first set of Germanic languages we analyze, we will show that a lower bound estimate obtained in terms of these distances is sharp, under a conditional assumption, which we discuss more in detail in Sect. 2.3.

The Euclidean distances of the flattening matrices from the corresponding determinantal varieties can be computed using the Eckart–Young theorem, as in Example 2.3 of [13] and [35].

The Eckart–Young theorem describes a low-rank approximation problem, namely minimizing the Euclidean distance $\|M - M'\|$ between a given $n \times m$ matrix M, seen as a vector in $\mathbb{R}^{nm}$, and an $n \times m$ matrix M' with $\operatorname{rank}(M') \leq k$, for a given $k \leq n \leq m$. One considers the singular value decomposition $M = U\Sigma V$ where Σ is an $n \times m$ diagonal matrix $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, and where U and V are, respectively $n \times n$ and $m \times m$ orthogonal matrices. Then the minimum of the distance $\|M - M'\|$ is realized by $M' = U\Sigma' V$ where $\Sigma' = \operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$ with the distance given by

$$\min_{M'} \|M - M'\| = \left(\sum_{i=k+1}^{n} \sigma_i^2 \right)^{1/2}.$$

This can equivalently be stated as the fact that the minimum distance between a given $n \times m$ matrix M and the determinantal variety $\mathcal{D}_k(n, m)$ of $n \times m$ matrices of rank $\leq k$ is given by

$$\operatorname{dist}(M, \mathcal{D}_k(n, m)) = \|(\sigma_{k+1}, \ldots, \sigma_n)\|, \tag{2.5}$$

where the σ_i are the singular values of M. The point M' realizing the minimum is unique iff $\sigma_{k+1} \neq \sigma_k$, with k the rank [31].

2.3 Conditional Cases and Distance Estimates

In the specific examples we discuss below, we usually consider a list of pre-selected candidate trees, obtained via the use of the PHYLIP package and among them we test for the most reliable candidate using the algebro-geometric methods discussed here. Unlike the case where the search happens over all possible interpolating binary trees, in these cases the pre-selected tree tend to all agree on certain proximity assignments of some of the leaves. For example, in the first set of Germanic languages that we discuss below, all the candidate trees agree on the proximity of Dutch and German and on the proximity of Icelandic and Faroese, though they disagree in the relative placements of these subtrees with respect to the other languages in the set. This agreement among the candidate trees results in two of the flattening matrices being common to all of the candidates.

In a situation like this one it is reasonable to consider a "conditional case" where we assume that the incidence condition that these common flattenings lie on the respective determinantal varieties already holds. We then aim at identifying the best candidate tree among those with these constraints already assumed.

We outline more precisely the reasoning behind the kind of estimation we are going to perform. We have a preselected small list of candidate trees T_i, $i = 1, \ldots, N$ and we assume that one of them is the correct phylogenetic tree. This assumption means that the point P given by the boundary distribution of i.i.d. variables that evolved according to a Markov model on this tree will lie on its phylogenetic variety. Thus, there is a T_{true} among the T_i for $i = 1, \ldots, N$ such that $P \in V_{T_{\text{true}}}$. If we also assume (as will be the case in specific examples we consider) that

all the phylogenetic varieties V_{T_i} are intersections of the form $V_{T_i} = W \cap V_i$, where W is common to all the T_i while the other varieties V_i depend on the tree T_i, then this assumption together with the previous one then gives $P \in V_{T_{\text{true}}} = W \cap V_{\text{true}}$ so necessarily $P \in W$. Thus, in this case the question about which of the varieties V_{T_i} the point P lies on is reduced to the question of which of the V_i the point lies on, as it will lie on W anyway. This would imply that it would suffice to check the Euclidean distances between P and the V_i.

However, because of possible noise in the data and other effects such as possible small discrepancies from the Markov hypothesis for syntactic parameters, we will in general have only a close proximity of P to the variety V_{T_i} of the correct phylogenetic tree, rather than exact incidence. We can account for possible small discrepancies by assuming that there is a sufficiently small $\epsilon > 0$ such that $P \in \mathcal{U}_\epsilon(V_{T_{\text{true}}})$, where T_{true} is correct phylogenetic tree and $V_{T_{\text{true}}} = V_{\text{true}} \cap W$, and $\mathcal{U}_\epsilon(V_{T_{\text{true}}})$ is an ϵ-tubular neighborhood of $V_{T_{\text{true}}}$ inside the ambient Euclidean space. With only this proximity estimate available, one can no longer necessarily relate which T_i realizes the minimum among the distances $\text{dist}(P, V_i)$ or the minimum among the $\text{dist}(P, V_i \cap W)$, as one could now have a situation where $\text{dist}(P, V_1 \cap W) < \text{dist}(P, V_2 \cap W)$ while $\text{dist}(P, V_2) < \text{dist}(P, V_1)$.

Nonetheless, if we compute the minimum Euclidean distances $\text{dist}(P, V_i)$, instead of directly obtaining the minimum among the distances $\text{dist}(P, W \cap V_i)$, this will provide a lower bound on the Euclidean distance $\text{dist}(P, V_{T_{\text{true}}})$. Indeed, we can simply obtain an estimate using the fact that the lower bound $\text{dist}(P, V \cap W) \geq \max\{\text{dist}(P, V), \text{dist}(P, W)\}$, for two subvarieties V, W in the same ambient space. Since this is only a lower bound, which is in general not expected to be sharp, one can at best hope to use this estimate to exclude candidates for which the computed $\max\{\text{dist}(P, V), \text{dist}(P, W)\}$ is large (within the set of given candidates), while a small value of this maximum will not necessarily imply that the corresponding candidate is optimal as $\text{dist}(P, V \cap W)$ could easily be significantly larger. We see however that in many cases this lower bound suffices to exclude most candidates hence it provides a useful estimate.

A more general theoretical discussion of these estimation methods and their range of validity, compared to other phylogenetic invariants and tree reconstruction algorithms (such as discussed in [8,14,45]) will be discussed elsewhere, separately from the present application, since they are not restricted to the specific linguistic setting considered here.

2.4 Limits of Applicability to Syntax

One of the purposes of this paper is also to better understand the limits of the applicability of these phylogenetic models to syntactic data. One of the main assumptions that need to be more carefully questioned is treating syntactic parameters as i.i.d. random variables evolving under the same Markov model on the tree. We know that there are relations between syntactic parameters. While the complete structure of the relations is not known, and is in fact one of the crucial questions in the field, one can detect the presence of relations through various computational methods applied to the available syntactic data.

In [30] and [46], a quantitative test was devised, aimed at measuring how the distribution of syntactic parameters over a group of languages differs from the result of i.i.d. random variables. Using coding theory, one associates a binary code to the set of syntactic parameters of a given group of languages and computes the position of the resulting code in the space of code parameters (the relative rate of the code and its relative minimum distance). If the distribution of the syntactic features across languages were the effect of an evolution of identically distributed independent random variables, one would expect to find the code points in the region of the space of code parameters populated by random codes in the Shannon random code ensembles, that is, in the region below the Gilbert–Varshamov curve. However, what one finds (see [46]) is the presence of many outliers that are not only above the Gilbert–Varshamov curve, but even above the symptotic bound and the Plotkin bound. This provides quantitative evidence for the fact that the evolutionary process that leads to the boundary distribution P of code parameters may differ significantly from the hypothesis of the phylogenetic model.

In [38] it was shown, using Kanerva networks, that different syntactic parameters in the SSWL database have different degrees of recoverability, which can be seen as another numerical indicator of the presence of relations, with

parameters with lower recoverability counting as closer to being truly independent variables and those with higher recoverability seen as dependent variables. One possible modification of the evolutionary model on the phylogenetic tree may then be obtained by computing the observed distribution P at the leaves, by introducing different weights for the different parameters, which depend on the recoverability factor, so that parameters that are more likely to be independent variables would weight more in determining the boundary distribution and parameters that have higher recoverability, and are therefore considered dependent variables, would contribute less to determining P.

A further issue worth mentioning, though we will not discuss it in this paper, is whether the hypothesis that the evolutionary dynamics happens on a tree is the best model. There are more general phylogenetic reconstruction techniques based on graphs that are not trees, see [19] and the algebro-geometric models in [9]. It was shown in [41] that the persistent topology of the SSWL data of some language families (the Indo-European) contain non-trivial persistent generators of the H_1 homology group. While the persistent generators of H_0 appear to be related to the structure of a candidate phylogenetic tree, the presence of a persistent H_1 points to the presence of loops, hence to graphs that are not trees. Persistent generators of the H_1 are also visible in the Longobardi data. This is further discussed in [42].

We discuss some possible modifications of the evolutionary Markov model on the tree in the last section of the paper.

3 Phylogenetic Algebraic Varieties of the Germanic Language Family

As discussed in the Introduction, we first analyze the phylogenetic tree for the set of Germanic languages $\mathcal{S}_1(G)$: Dutch, German, English, Faroese, Icelandic, and Swedish.

These six languages are mapped with different levels of accuracy in the SSWL database: we have Dutch (100%), German (75%), English (75%), Faroese (62%), Icelandic (62%), Swedish (75%). There are 90 syntactic variables that are completely mapped for all of these six languages: the list is reported in Appendix A. We will use only these 90 variables for the analysis carried out here.

We then consider the set $\mathcal{S}_2(G)$ consisting of seven Germanic languages: Norwegian, Danish, Icelandic, German, English, Gothic, Old English. These are chosen so that they are covered by both the SSWL database [49] and the new data of Longobardi [24], and so that they contain some ancient languages, in addition to modern languages situated on both the West and the North Germanic branches. In this way we can test both the effect of using different syntactic data and the effect of including ancient languages and their relation to problem of the location of the root vertex mentioned above.

The Germanic languages in the set $\mathcal{S}_2(G)$ have a total of 68 SSWL variables that are completely mapped for all the seven languages in the set. This is significantly smaller than the 90 variables used for the set $\mathcal{S}_1(G)$. This does not depend on the languages being poorly mapped: the levels of accuracy are comparable with the previous set with Danish (76%), Norwegian (75%), German (75%), English (75%), Old English (75%) Icelandic (62%), Gothic (62%). However, the regions of the overall 115 SSWL variables that are mapped is less uniform across this set of languages creating a smaller overlap. The set of completely mapped SSWL variables for this set of languages is reported in Appendix B.

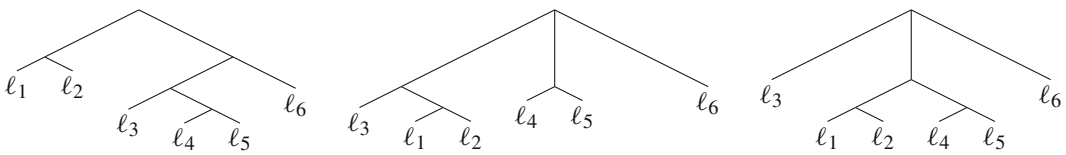
3.1 Candidate PHYLIP Trees

When using the full but incomplete data for the six Germanic languages in $\mathcal{S}_1(G)$, we obtain with PHYLIP a list of six candidate phylogenetic trees, respectively given (in bracket notation) by

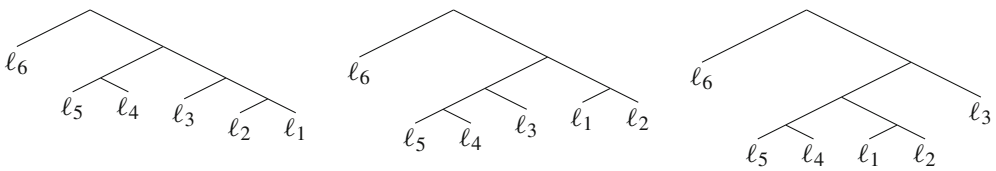
$$\texttt{pars1} = ((\ell_1, \ell_2), (\ell_3, (\ell_4, \ell_5)), \ell_6)$$
$$\texttt{pars2} = ((\ell_3, (\ell_1, \ell_2)), (\ell_4, \ell_5), \ell_6)$$
$$\texttt{pars3} = (\ell_3, ((\ell_1, \ell_2), (\ell_4, \ell_5)), \ell_6)$$
$$\texttt{bnb1} = (\ell_6, ((\ell_5, \ell_4), (\ell_3, (\ell_2, \ell_1))))$$
$$\texttt{bnb2} = (\ell_6, (((\ell_5, \ell_4), \ell_3), (\ell_1, \ell_2)))$$
$$\texttt{bnb3} = (\ell_6, (((\ell_5, \ell_4), (\ell_1, \ell_2)), \ell_3))$$

where ℓ_1 = Dutch, ℓ_2 = German, ℓ_3 = English, ℓ_4 = Faroese, ℓ_5 = Icelandic, ℓ_6 = Swedish. The Newick representation of binary trees used by PHYLIP lists the leaves in the order specified by the choice of a planar embedding of the tree, with brackets and commas indicating the joining together of branches. In the rest of the paper, for convenience, we will spell out explicitly the form of the tree graphically, rather than writing them in the Newick bracket notation. In the case of the trees listed here we obtain the following.

The trees `pars1`, `pars2`, and `pars3` given above in the Newick representation have the form



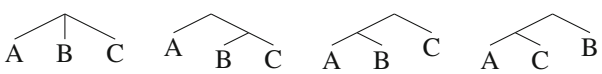
Note that `pars1` is a binary tree, while `pars2` and `pars3` are not binary trees. We will discuss how to resolve the non-binary structure. The remaining trees `bnb1`, `bnb2`, and `bnb3` are binary trees of the form



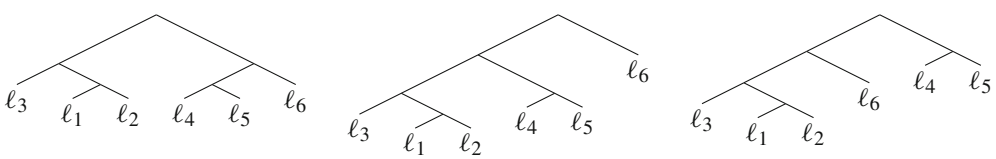
Note how all of these candidate trees agree on the proximity of Dutch and German (ℓ_1 and ℓ_2) and of Faroese and Icelandic (ℓ_4 and ℓ_5), while they differ in the relative placement of these two pairs with respect to one another and with respect to the two remaining languages, English and Swedish.

In phylogenetic linguistics the presence of a non-binary tree denotes an ambiguity, which should eventually be resolved into one of its possible binary splits. As shown in [15], the phylogenetic algebraic variety of a non-binary tree can be seen as the intersection of the phylogenetic algebraic varieties of all of its possible binary splits. Thus, the phylogenetic ideal (for the binary Jukes-Cantor model) is generated by all the 3×3 minors of all the flattening matrices of all the binary splits of the given non-binary tree. Being the intersection of the varieties defined by each of the binary splits corresponds exactly to the notion of ambiguity mentioned above.

The resolution of a non-binary structure of the type shown in `pars2` and `pars3` is obtained by replacing the first tree below with the different possibilities given by its three possible binary splits that follow:

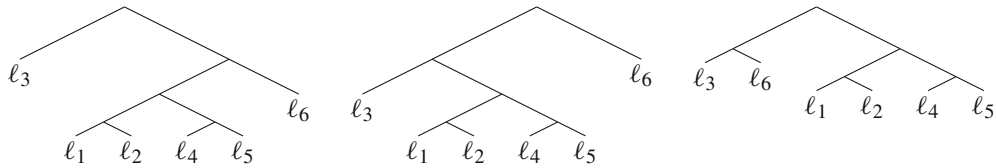


Thus, for the tree `pars2` we obtain the three binary trees



Note, however, that these three binary trees are equivalent up to a shift in the position of the root, which however does not affect the phylogenetic invariants, see [1] and Proposition 2.16 in [5]. Thus, we need only consider one

of them for the purpose of computing the generators of the phylogenetic ideal. For the tree pars3 we obtain the three binary trees



Again these three binary trees only differ by a shift of the position of the root, which does not affect the computation of the phylogenetic invariants, hence we need only consider one of them for that purpose. Notice, moreover, that the binary tree bnb1 is the same as the second binary tree for pars2. Also the tree bnb2 has the same topology as the tree pars1, up to a shift in the position of the root, which does not affect the phylogenetic invariants. Similarly, the tree bnb3 is the same as the second binary tree of pars3.

All of the binary trees considered here have three internal edges, hence all of them have three flattenings $\mathrm{Flat}_{e,T}(P)$ of the boundary distribution $P = (p_{i_1,\dots,i_6})$.

- The flattenings for pars1 are given by a 4×16 matrix $\mathrm{Flat}_{e_1,\mathtt{pars1}}(P)$, an 8×8 matrix $\mathrm{Flat}_{e_2,\mathtt{pars1}}(P)$ and a 16×4 matrix $\mathrm{Flat}_{e_3,\mathtt{pars1}}(P)$. These correspond to the separating the leaves into two components when deleting the internal edge e_i according to

  $e_1 : \{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5, \ell_6\}$

  $e_2 : \{\ell_1, \ell_2, \ell_6\} \cup \{\ell_3, \ell_4, \ell_5\}$

  $e_3 : \{\ell_1, \ell_2, \ell_3, \ell_6\} \cup \{\ell_4, \ell_5\}$.

- The flattenings for any of the three binary trees for pars2 are also given by a 4×16 matrix $\mathrm{Flat}_{e_1,\mathtt{pars2}}(P)$, an 8×8 matrix $\mathrm{Flat}_{e_2,\mathtt{pars2}}(P)$ and a 16×4 matrix $\mathrm{Flat}_{e_3,\mathtt{pars2}}(P)$, which in this case correspond to the subdivisions

  $e_1 : \{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5, \ell_6\}$

  $e_2 : \{\ell_1, \ell_2, \ell_3\} \cup \{\ell_4, \ell_5, \ell_6\}$

  $e_3 : \{\ell_1, \ell_2, \ell_3, \ell_6\} \cup \{\ell_4, \ell_5\}$,

  which only differ from the previous case in the e_2 flattening.

- The flattenings for any of the three binary trees for pars3 are given by a 4×16 matrix $\mathrm{Flat}_{e_1,\mathtt{pars3}}(P)$, a 16×4 matrix $\mathrm{Flat}_{e_2,\mathtt{pars3}}(P)$ and a 16×4 matrix $\mathrm{Flat}_{e_3,\mathtt{pars3}}(P)$, which correspond to the subdivisions

  $e_1 : \{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5, \ell_6\}$

  $e_2 : \{\ell_1, \ell_2, \ell_3, \ell_6\} \cup \{\ell_4, \ell_5\}$

  $e_3 : \{\ell_1, \ell_2, \ell_4, \ell_5\} \cup \{\ell_3, \ell_6\}$.

- The bnb1 tree is the same as one of binary trees for pars2, hence their flattenings are also the same.
- The flattenings for bnb2 are the same as the flattening of pars1, since the two tree differ only by a shift in the position of the root vertex.
- The bnb3 tree is the same as one of binary trees for pars3, hence their flattenings are also the same.

Thus, in order to compare the phylogenetic invariants of these various trees, we need to compute the 3×3 minors of the matrices $\mathrm{Flat}_{e,T}(P)$ for the splits $\{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5, \ell_6\}$, $\{\ell_1, \ell_2, \ell_6\} \cup \{\ell_3, \ell_4, \ell_5\}$, $\{\ell_1, \ell_2, \ell_3, \ell_6\} \cup \{\ell_4, \ell_5\}$, $\{\ell_1, \ell_2, \ell_3\} \cup \{\ell_4, \ell_5, \ell_6\}$, $\{\ell_1, \ell_2, \ell_4, \ell_5\} \cup \{\ell_3, \ell_6\}$. We will compute these in the next subsection.

3.2 Flattenings

As discussed above, there are five matrices $\mathrm{Flat}_{e,T}(P)$ that occur in the computation of the phylogenetic ideals of the candidate phylogenetic trees listed above. In fact, we do not need to compute all of them, as some occur in all

the trees, hence do not contribute to distinguishing between them. This corresponds to the observation we already made above, that all the candidate trees agree on the proximity of ℓ_1 and ℓ_2 and of ℓ_4 and ℓ_5.

To simplify keeping track visually of which flattening is being considered, we replace here the edge notation e of the flattening matrices $\mathrm{Flat}_{e,T}(P)$ with the explicit splitting of the leaves of T that corresponds to the edge e. Thus, for example, instead of writing $\mathrm{Flat}_{e_1,\texttt{pars1}}(P)$ we write $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P)$. This notation has the advantage that, when the same flattening matrix (with the same subdivision of leaves) occurs in different trees, this will be immediately evident from the notation. We will continue to use the more concise notation $\mathrm{Flat}_{e,T}(P)$ when more convenient.

- The 4×16 matrix $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P)$, contributes to the phylogenetic ideals of all the trees, hence it will not help discriminate between them.
- The same is true about the 16×4 matrix $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P)$.
- The 8×8 matrix $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$ contributes to the phylogenetic invariants of `pars1` and `bnb2`. It is given by

$$\begin{pmatrix}
p_{000000} & p_{000100} & p_{001000} & p_{001100} & p_{000010} & p_{000110} & p_{001010} & p_{001110} \\
p_{010000} & p_{010100} & p_{011000} & p_{011100} & p_{010010} & p_{010110} & p_{011010} & p_{011110} \\
p_{100000} & p_{100100} & p_{101000} & p_{101100} & p_{100010} & p_{100110} & p_{101010} & p_{101110} \\
p_{110000} & p_{110100} & p_{111000} & p_{111100} & p_{110010} & p_{110110} & p_{111010} & p_{111110} \\
p_{000001} & p_{000101} & p_{001001} & p_{001101} & p_{000011} & p_{000111} & p_{001011} & p_{001111} \\
p_{010001} & p_{010101} & p_{011001} & p_{011101} & p_{010011} & p_{010111} & p_{011011} & p_{011111} \\
p_{100001} & p_{100101} & p_{101001} & p_{101101} & p_{100011} & p_{100111} & p_{101011} & p_{101111} \\
p_{110001} & p_{110101} & p_{111001} & p_{111101} & p_{110011} & p_{110111} & p_{111011} & p_{111111}
\end{pmatrix}$$

- The 8×8 matrix $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$ contributes to the phylogenetic invariants of `pars2` and `bnb1` and it is given by

$$\begin{pmatrix}
p_{000000} & p_{000010} & p_{000100} & p_{000110} & p_{000001} & p_{000011} & p_{000101} & p_{000111} \\
p_{010000} & p_{010010} & p_{010100} & p_{010110} & p_{010001} & p_{010011} & p_{010101} & p_{010111} \\
p_{100000} & p_{100010} & p_{100100} & p_{100110} & p_{100001} & p_{100011} & p_{100101} & p_{100111} \\
p_{110000} & p_{110010} & p_{110100} & p_{110110} & p_{110001} & p_{110011} & p_{110101} & p_{110111} \\
p_{001000} & p_{001010} & p_{001100} & p_{001110} & p_{001001} & p_{001011} & p_{001101} & p_{001111} \\
p_{011000} & p_{011010} & p_{011100} & p_{011110} & p_{011001} & p_{011011} & p_{011101} & p_{011111} \\
p_{101000} & p_{101010} & p_{101100} & p_{101110} & p_{101001} & p_{101011} & p_{101101} & p_{101111} \\
p_{111000} & p_{111010} & p_{111100} & p_{111110} & p_{111001} & p_{111011} & p_{111101} & p_{111111}
\end{pmatrix}$$

- The 16×4 matrix $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$ contributes to the phylogenetic invariants of `pars3` and `bnb3` and is given by

$$\begin{pmatrix}
p_{000000} & p_{000001} & p_{001000} & p_{001001} \\
p_{010000} & p_{010001} & p_{011000} & p_{011001} \\
p_{100000} & p_{100001} & p_{101000} & p_{101001} \\
p_{110000} & p_{110001} & p_{111000} & p_{111001} \\
p_{000010} & p_{000011} & p_{001010} & p_{001011} \\
p_{010010} & p_{010011} & p_{011010} & p_{011011} \\
p_{100010} & p_{100011} & p_{101010} & p_{101011} \\
p_{110010} & p_{110011} & p_{111010} & p_{111011} \\
p_{000100} & p_{000101} & p_{001100} & p_{001101} \\
p_{010100} & p_{010101} & p_{011100} & p_{011101} \\
p_{100100} & p_{100101} & p_{101100} & p_{101101} \\
p_{110100} & p_{110101} & p_{111100} & p_{111101} \\
p_{000110} & p_{000111} & p_{001110} & p_{001111} \\
p_{010110} & p_{010111} & p_{011110} & p_{011111} \\
p_{100110} & p_{100111} & p_{101110} & p_{101111} \\
p_{110110} & p_{110111} & p_{111110} & p_{111111}
\end{pmatrix}$$

3.3 Boundary Distribution and Phylogenetic Invariants

Next we compute the boundary distribution $P = (p_{i_1,\ldots,i_6})$ of the syntactic variables. We use only the 90 completely mapped syntactic variables, for which we find occurrences

$$
\begin{array}{llll}
n_{110111} = 3 & n_{000011} = 1 & n_{000010} = 4 & n_{000000} = 40 \\
n_{110000} = 2 & n_{001110} = 1 & n_{000100} = 2 & n_{111111} = 22 \\
n_{111110} = 1 & n_{000110} = 1 & n_{111101} = 3 & n_{100000} = 2 \\
n_{010000} = 1 & n_{111001} = 2 & n_{110110} = 1 & n_{010111} = 1 \\
n_{001000} = 2 & n_{000111} = 1 &
\end{array}
$$

while all the remaining cases do not occur, $n_{i_1,\ldots,i_6} = 0$ for $(i_1, \ldots, i_n)$ not in the above list.

With the boundary distribution determined by the occurrences above the three matrices of $F1 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$, $F_2 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$, and $F_3 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$ are, respectively, given by

$$
F_1 = \begin{pmatrix}
\frac{4}{9} & \frac{1}{45} & \frac{1}{45} & 0 & \frac{2}{45} & \frac{1}{90} & 0 & \frac{1}{90} \\
\frac{1}{90} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & 0 & 0 & \frac{1}{90} & 0 & \frac{1}{90} \\
0 & 0 & 0 & 0 & \frac{1}{90} & \frac{1}{90} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{90} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{45} & \frac{1}{30} & 0 & \frac{1}{30} & 0 & \frac{11}{45}
\end{pmatrix}
$$

$$
F_2 = \begin{pmatrix}
\frac{4}{9} & \frac{2}{45} & \frac{1}{45} & \frac{1}{90} & 0 & \frac{1}{90} & 0 & \frac{1}{90} \\
\frac{1}{90} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{90} \\
\frac{1}{45} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & \frac{1}{90} & 0 & 0 & 0 & \frac{1}{30} \\
\frac{1}{45} & 0 & 0 & \frac{1}{90} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{90} & \frac{1}{45} & 0 & \frac{1}{30} & \frac{11}{45}
\end{pmatrix}
$$

$$F_3 = \begin{pmatrix} \frac{4}{9} & 0 & \frac{1}{45} & 0 \\ \frac{1}{90} & 0 & 0 & 0 \\ \frac{1}{45} & 0 & 0 & 0 \\ \frac{1}{45} & 0 & 0 & \frac{1}{45} \\ \frac{2}{45} & \frac{1}{90} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{45} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{30} \\ \frac{1}{90} & \frac{1}{90} & \frac{1}{90} & 0 \\ 0 & \frac{1}{90} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{90} & \frac{1}{30} & \frac{1}{90} & \frac{11}{45} \end{pmatrix}$$

3.4 Phylogenetic Invariants

As we discussed above, the splits

$\{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5, \ell_6\}$ and $\{\ell_1, \ell_2, \ell_3, \ell_6\} \cup \{\ell_4, \ell_5\}$

occur in all the candidate trees, hence the minors coming from the flattening matrices

$\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P)$ and $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P)$

do not discriminate between the given candidates (preselected by PHYLIP). Thus it is reasonable to proceed by assuming that the condition that these two flattenings lie on the corresponding determinantal varieties is satisfied and only discriminate between the candidate trees on the basis of the position of the remaining flattenings. There is only one additional flattening involved in each tree, once these common ones are excluded. Thus, we estimate the phylogenetic invariants by evaluating the 3×3 minors of the remaining flattening matrix for each of the trees, using both the ℓ^∞ and the ℓ^1 norm. We obtain the following:

(1) For the tree $T_1 = \texttt{pars1}$ (and equivalently `bnb2`) we have

$$\|\phi_{T_1}(P)\|_{\ell^\infty} = \max_{\phi \in \mathcal{D}^{(3)}_{T_1}} |\phi(P)| = \frac{22}{18225}$$

$$\|\phi_{T_1}(P)\|_{\ell^1} = \sum_{\phi \in \mathcal{D}^{(3)}_{T_1}} |\phi(P)| = \frac{3707}{364500}$$

(2) For the tree $T_2 = \texttt{pars2}$ (equivalently `bnb1`) we have

$$\|\phi_{T_2}(P)\|_{\ell^\infty} = \max_{\phi \in \mathcal{D}^{(3)}_{T_2}} |\phi(P)| = \frac{419}{364500}$$

$$\|\phi_{T_2}(P)\|_{\ell^1} = \sum_{\phi \in \mathcal{D}^{(3)}_{T_2}} |\phi(P)| = \frac{2719}{364500}$$

(3) For the tree $T_3 = \texttt{pars3}$ (and equivalently bnb3) we have

$$\|\phi_{T_3}(P)\|_{\ell^\infty} = \max_{\phi \in \mathcal{D}^{(3)}_{T_3}} |\phi(P)| = \frac{22}{18225}$$

$$\|\phi_{T_3}(P)\|_{\ell^1} = \sum_{\phi \in \mathcal{D}^{(3)}_{T_3}} |\phi(P)| = \frac{949}{91125}$$

Thus, in terms of the evaluation of the phylogenetic invariants, the binary trees of pars2 and the binary tree bnb1 are favored over the other possibilities. (We discuss the position of the root vertex below.) Note that the ℓ^∞ norm does not distinguish between the other two remaining candidates and only singles out the preferred candidate pars2. We compute the Euclidean distance function in Sect. 3.7.

3.5 The Problem with the Root Vertex

As we have seen above, the computation of the phylogenetic invariants helps selecting between different candidate tree topologies. However, the phylogenetic invariants by themselves are insensitive to changing the position of the root in binary trees with the same topology. In terms of phylogenetic inference about linguistics, however, it is important to locate more precisely where the root vertex should be. In the case of languages belonging to a subfamily of the Indo-European languages this can be done, as in the example we discussed in [47], by introducing the data of some of the ancient languages in the same subfamily as a new leaf of the tree, that will help locating more precisely the root vertex of the original tree based on the modern languages. For language families for which there are no data of ancient languages available, however, this kind of phylogenetic analysis will only identify a tree topology as an unrooted binary tree. We will return to this point in the following section, where we analyze the set $\mathcal{S}_2(G)$ which includes two ancient languages.

Note that when one or more ancient languages are included in the data (as in the second case of the Germanic languages, or the Romance languages discussed here) that suffices to constrain the position of the root vertex, while in other cases like the example discussed here, additional independent information is needed.

3.6 Varieties

In the discussion above we reduced the question of distinguishing between the candidate trees to an evaluation of the phylogenetic invariants coming from the 3×3 minors of one of the three matrices $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$, $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$, and $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$. In the first two cases, the phylogenetic ideal defines the 28-dimensional determinantal variety of all 8×8 matrices of rank at most two, while in the third case the phylogenetic ideal defines the 36-dimensional determinantal variety of all 16×4 matrices of rank at most two, [7]. These are not the actual phylogenetic varieties associated to the candidate trees, which are further cut out by the remaining equations coming from the 3×3 minors of the other flattenings $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P)$, and $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P)$. The varieties associated to each individual tree are intersections of three different determinantal varieties inside a common ambient space $\mathbb{A}^{2^6}$. Since all the polynomials defining the phylogenetic ideals are homogeneous, they can also be considered as projective varieties in the ambient projective space $\mathbb{P}^{2^6-1}$.

In the case of the trees considered here, two of the three determinantal varieties stay the same, since the flattenings $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P)$, and $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P)$ are common to all candidate trees, while the third component varies among the three choices determined by the flattenings $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$, $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$, and $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$.

In general, let $\mathcal{D}_r(n,m)$ denote the determinantal variety of $n \times m$ matrices of rank $\leq r$. As an affine subvariety in $\mathbb{A}^{nm}$ it has dimension $r(n+m-r)$. It will be convenient to consider $\mathcal{D}_r(n,m)$ as a projective subvariety of $\mathbb{P}^{nm-1}$, though we will maintain the same notation. In the case $r = 1$, the determinantal variety $\mathcal{D}_1(n,m)$ is the Segre variety

$\mathcal{S}(n,m)$ given by the embedding $\mathbb{P}^{n-1} \times \mathbb{P}^{m-1} \hookrightarrow \mathbb{P}^{nm-1}$ realized by the Segre map $(x_i, y_j) \mapsto (u_{ij} = x_i y_j)$. In the case $r = 2$ the determinantal variety $\mathcal{D}_2(n,m)$ is the secant variety of lines (chord variety) $\mathrm{Sec}(\mathcal{S}(n,m))$ of the Segre variety $\mathcal{S}(n,m)$, see §9 of [20].

Thus, we obtain the following simple geometric description of the three cases considered above:

- $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$ (tree topology of `pars1` and `bnb2`): the relevant variety is the secant variety $\mathrm{Sec}(\mathcal{S}(8,8))$ of the Segre variety $\mathcal{S}(8,8) = \mathbb{P}^7 \times \mathbb{P}^7$, embedded in $\mathbb{P}^{63}$ via the Segre embedding $u_{i_1,\dots,i_6} = x_{i_1,i_2,i_6} y_{i_3,i_4,i_5}$.
- $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$ (tree topology of `pars2` and `bnb1`): the relevant variety is again $\mathrm{Sec}(\mathcal{S}(8,8))$, where $\mathcal{S}(8,8)$ is embedded in $\mathbb{P}^{63}$ via $u_{i_1,\dots,i_6} = x_{i_1,i_2,i_3} y_{i_4,i_5,i_6}$.
- $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$ (tree topology of `pars3` and `bnb3`): the relevant variety is the secant variety $\mathrm{Sec}(\mathcal{S}(16,4))$ of the Segre variety $\mathcal{S}(16,4) = \mathbb{P}^{15} \times \mathbb{P}^3$, embedded in $\mathbb{P}^{63}$ via the Segre embedding $u_{i_1,\dots,i_6} = x_{i_1,i_2,i_4,i_5} y_{i_3,i_6}$.

The evaluation of the phylogenetic invariants at the boundary distribution determined by the SSWL data selects the second choice, $\mathrm{Sec}(\mathcal{S}(8,8))$ with the Segre embedding $u_{i_1,\dots,i_6} = x_{i_1,i_2,i_3} y_{i_4,i_5,i_6}$.

As a general procedure, given a subfamily of languages, $\{\ell_1,\dots,\ell_n\}$ and a set of candidate phylogenetic trees $T_1,\dots,T_m$ produced by computational methods from the syntactic variables of these n languages, one can construct with the method above a collection $Y_1,\dots,Y_m$ of algebraic varieties, where each Y_k associated to the tree T_k is obtained by considering the determinantal varieties associated to all those flattenings $\mathrm{Flat}_{e,T_k}(P)$ of T_k that are not common to all the other trees T_j.

The test for selecting one of the candidate trees, given the boundary distribution $P = (p_{i_1,\dots,i_n})$ of the syntactic variables, is then to estimate which of the varieties Y_k the point P is closest to, where a suitable test of closeness is used, for instance through the Euclidean distance function. Assuming that this procedure does not result in ambiguities (that is, that there is a unique closest Y_k to the given distribution P), then this method selects a best candidate T among the m trees T_k. It also selects an associated algebraic variety $Y = Y(T)$, which is larger than the usual phylogenetic algebraic variety X_T of T, since we have neglected flattenings that occur simultaneously in all the m candidate trees T_k.

3.7 The Euclidean Distance

According to the discussion of the previous subsection, on the geometry of the varieties involved in distinguishing between the candidate trees, we compute here

- the Euclidean distance of the point $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$ and the determinantal variety $\mathcal{D}_2(8,8) = \mathrm{Sec}(\mathcal{S}(8,8))$,
- the Euclidean distance of the point $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$ from the same determinantal variety $\mathcal{D}_2(8,8) = \mathrm{Sec}(\mathcal{S}(8,8))$,
- the Euclidean distance of the point $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$ from the determinantal variety $\mathcal{D}_2(16,4) = \mathrm{Sec}(\mathcal{S}(16,4))$.

Using the Eckart-Young theorem, we compute these distances using the singular values of these three matrices. These are given by

$$\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)) \sim$$
$$\mathrm{diag}(0.44940, 0.25001, 0.19237 \times 10^{-1}, 0.96007 \times 10^{-2}, 0.21595 \times 10^{-2}, 0.88079 \times 10^{-3}, 4.6239 \times 10^{-19}, 0)$$
$$\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)) \sim$$
$$\mathrm{diag}(0.44956, 0.25018, 0.14729 \times 10^{-1}, 0.44229 \times 10^{-2}, 0.27802 \times 10^{-2}, 0.24881 \times 10^{-17}, 0)$$
$$\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)) \sim$$
$$\mathrm{diag}(0.44939, 0.24994, 0.20625 \times 10^{-1}, 0.94442 \times 10^{-2}).$$

Using (2.5) we then obtain

$$\operatorname{dist}(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P), \mathrm{Sec}(\mathcal{S}(8,8)))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.46768 \times 10^{-3}$$
$$\operatorname{dist}(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P), \mathrm{Sec}(\mathcal{S}(8,8)))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.24424 \times 10^{-3}$$
$$\operatorname{dist}(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P), \mathrm{Sec}(\mathcal{S}(16,4)))^2 = \sigma_3^2 + \sigma_4^2 = 0.51457 \times 10^{-3}$$

The second Euclidean distance is the smallest, hence this more reliable distance test again favors the binary trees of `pars2` and the binary tree `bnb1`.

The computation of these Euclidean distances provides a selection between the candidate trees in the following way. The first distance measures how far the point determined by the data (in the form of the boundary distribution P and the flattening matrix $F_1(P)$) is from the determinantal variety $\mathcal{D}_2(8,8)$ determined by the tree `pars1`. The second distance measures how far the point determined by the data, through the flattening $F_2(P)$, is from the determinantal variety determined by the tree `pars2`, and the third distance measures how far the point, through the flattening $F_3(P)$ is from the determinantal variety $\mathcal{D}_2(16,4)$ determined by the tree `pars3`. Since as observed above the remaining flattenings of P occur in all trees and do not help distinguishing between them, it suffices to find the best matching condition between the three possibilities listed here, for which we select the one realizing the smallest Euclidean distance.

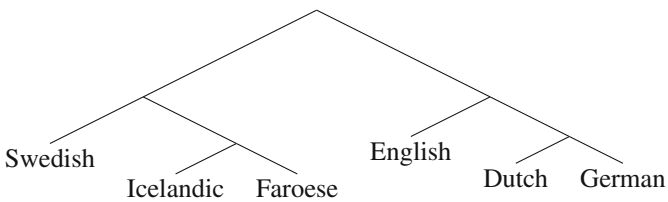
The Euclidean distances computed above provide lower bound estimates for the distances $\operatorname{dist}(P, V_{T_i})$. Even though these are just lower bounds, they do agree with the phylogenetic invariants test in the selection of the candidate trees. Heuristically, we can think of this as reflecting the fact that the determinantal varieties associated to the flattening matrices

$$\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P) \quad \text{and} \quad \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P)$$

that are common to all the tree candidates are not contributing in discriminating among the different T_i (though see the more precise discussion in Sect. 2.3 above).

3.8 The West/North Germanic Split from SSWL Data

Note that the tree topology selected in this way, which (up to the position of the root vertex) is equivalent to the tree

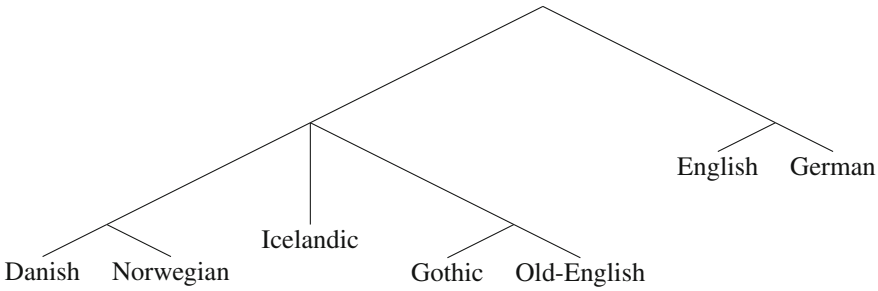


is also the generally acknowledged correct subdivision of the Germanic languages into the North Germanic and the West Germanic sub-branches. The North Germanic in turn splits into a sub-brach that contains Swedish (but also Danish which we have not included here) and another that contains Icelandic and Faroese (and also Norwegian, which we have not included, in order to keep the number of leaves more manageable). The West Germanic branch is split into the Anglo-Frisian sub-branch (of which here we are only considering English, but which should also contain Frisian) and the Netherlandic-Germanic branch that contains Dutch and German. Thus, the analysis through phylogenetic invariants and the estimate of the Euclidean distance have selected the correct tree topology among the candidates produced by the computational analysis of the SSWL data obtained with PHYLIP.

3.9 Longobardi Data and Phylogenetic Invariants of Germanic Languages

Now we analyze the set $\mathcal{S}_2(G)$ consisting of Norwegian, Danish, Icelandic, German, English, Gothic, and Old English, using the syntactic parameters collected in the new data of Longobardi [24].

The DNA parsimony algorithm of PHYLIP based solely on the new Longobardi data produces a single candidate phylogenetic tree for the set $\mathcal{S}_2(G)$ of Germanic languages, of the form

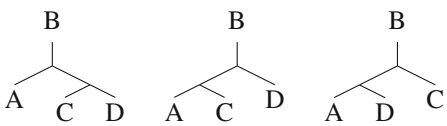


In fact, because of the presence of a vertex of higher valence in this tree, one should resolve it into the possible binary trees and compare the resulting candidates. Moreover, the placement of the ancient languages as "leaves" of the tree is an artifact, and needs to be resolved into the appropriate placement of the root of the binary trees.

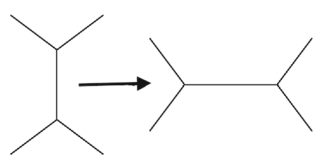
We see here that the fact that ancient languages are treated as leaves in the tree although they really are intermediate nodes creates some problems in the reconstruction provided by PHYLIP. In the PHYLIP tree above Gothic and Old English are grouped as nearby leaves in the tree, since the reconstruction correctly identifies the closer proximity of the two ancient languages with respect to the modern ones. However, this causes an error in the proposed tree topology when these are placed as two nearby leaves. The standard way of resolving the higher valence vertex, as discussed in the previous section, would maintain this problem. We propose here a simple method for avoiding this problem, via a simple topological move in the resulting trees that restores the role of these two languages as intermediate nodes of the tree (and suggests a position of the root vertex) while maintaining their relation to the rest of the tree.

In particular, this means that we are going to consider possible candidate trees of the following form, where we set $\ell_1 =$ Norwegian, $\ell_2 =$ Danish, $\ell_3 =$ Gothic, $\ell_4 =$ Old English, $\ell_5 =$ Icelandic, $\ell_6 =$ English, $\ell_7 =$ German.

We first visualize the trees obtained by resolving the triple vertex. To simplify the picture, let us write $A = \{\ell_1, \ell_2\}$ for the end of the tree containing this pair of adjacent leaves, and similarly for $B = \{\ell_3, \ell_4\}$, $C = \{\ell_5\}$, $D = \{\ell_6, \ell_7\}$, so that we can visualize the the three possible binary splits of the vertex in the PHYLIP tree as the trees



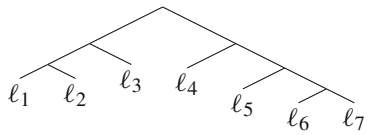
We then want to input the extra piece of information concerning the fact that the leaves in the set $B = \{\ell_3, \ell_4\}$ are not really leaves but inner vertices of the tree, whose proximity is describing the fact that they are in closer proximity to the root of the tree than the other leaves, rather than their proximity as leaves. We argue that this can be done effectively by introducing a simple *topological move* on these trees that achieves exactly this effect, while preserving the relation to the rest of the tree, namely the following operation:



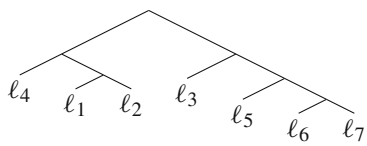
More explicitly, this means the following. Suppose that a configuration as in the left-hand-side appears in a candidate tree, where the two bottom leaves are ancient languages placed as nearby leaves of the tree, and the two top directions continue to other branches of the tree. One replaces it, without changing the rest of the tree, with the configuration on the right-hand-side. In this configuration, the two bottom leaves are still labelled by the same two ancient languages and the two top directions are still attached to the same other branches of the tree to which they were connected in the

left-hand-side. The configuration obtained in this way represents more correctly the role of the ancient languages, by assigning to each of them an internal vertex of the tree, the vertex to which the leaf is now attached. Note that on the right-hand-side there are two choices of how to place the labels in the two lower leaves: permuting the two lower leaves in the left-hand-side has no effect, but permuting them on the right-hand-side gives rise to two different tree candidates, both of which need to be taken into consideration. In a case like the present one, where these are the only two ancient languages in the tree, this also suggests that the root vertex should be placed in between these two points. Applying this operation produces the following list of candidate trees, with (1) and (2) derived from the first binary tree above, (3) and (4) from the second binary tree above and (5) and (6) from the third one. Note that each of these pairs corresponds to the two possible choices of labels in the right-hand-side, as mentioned above.

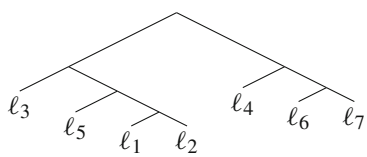
(1) The first candidate tree $T_1(G)$ has Icelandic (incorrectly) grouped together with the West Germanic (German, English) instead of the North Germanic (Norwegian, Danish) languages. The labels ℓ_3 and ℓ_4 should be thought of not as leaves but as intermediate vertices placed, respectively, above the $\{\ell_1, \ell_2\}$ subtree and above the $\{\ell_5, \ell_6, \ell_7\}$ subtree.



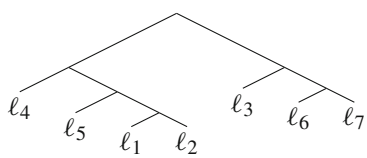
(2) The second candidate tree $T_2(G)$ has the same structure as the previous list (with the incorrect placement of Icelandic), but with the reversed placement of the two ancient languages ℓ_3 and ℓ_4, this time with Old English placed at the top of the North Germanic instead of the West Germanic subtree:



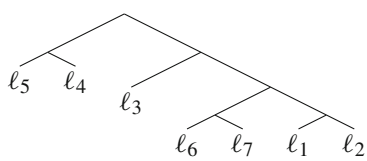
(3) The third candidate tree $T_3(G)$ has the correct placement of Icelandic in the North Germanic subtree, with Gothic above the North Germanic and Old English above the West Germanic subtrees:



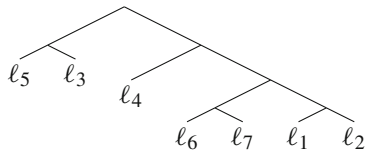
(4) The fourth candidate tree $T_4(G)$ also has the correct placement of Icelandic in the North Germanic subtree, now with Old English above the North Germanic and Gothic above the West Germanic subtrees:



(5) The fifth candidate incorrectly places the sets $\{\ell_1, \ell_2\}$ and $\{\ell_6, \ell_7\}$ in closer proximity and ℓ_5 in a separate branch away from the ancient languages $\{\ell_3, \ell_4\}$, placing ℓ_4 as the ancient language in closer proximity to ℓ_5:



(6) The sixth candidate tree also incorrectly places ℓ_5 as a separate branch and $\{\ell_1, \ell_2\}$ and $\{\ell_6, \ell_7\}$ in the same branch, while placing ℓ_3 as the ancient language in closer proximity to ℓ_5:



We first discuss the candidate trees (1)–(4) as these have a lot of common structure that simplifies a common analysis. We then show what changes for the last two cases.

When considering the new Longobardi data for the purpose of computing phylogenetic invariants, we need to eliminate from the list all those parameters that have value either 0 (undefined in the terminology of Longobardi's data table) or ? (unknown). The reason for eliminating not just the unknown parameters but also those rendered undefined by entailment relations lies in the fact that the result of [1] that we use for the computation of the phylogenetic invariants holds for a *binary* Jukes-Cantor model but not for a ternary one. Thus, we stick to only those parameters that are defined with binary values ± 1 in Longobardi's table, for all the languages $\ell_1, \dots, \ell_7$ in our list of Germanic languages. After the change of notation to binary form, obtained by replacing $1 \mapsto 1$ and $-1 \mapsto 0$, we obtain the following list of parameters

$\ell_1 = [1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0]$

$\ell_2 = [1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0]$

$\ell_3 = [1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0]$

$\ell_4 = [1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0]$

$\ell_5 = [1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0]$

$\ell_6 = [1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0]$

$\ell_7 = [1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0]$

Notice how one is left with a shorter list of only 42 parameters, where most of them have the same value for all the languages in this group. The only non-zero frequencies for binary vectors $(a_1, \dots, a_7) \in \mathbb{F}_2^7$ that arise in the boundary distribution at the leaves of the trees are

$$n_{1111111} = 12 \; n_{0000000} = 24 \; n_{1101111} = 1 \; n_{1111101} = 1$$
$$n_{1111100} = 1 \quad n_{1111011} = 1 \quad n_{1100111} = 1 \; n_{0011111} = 1$$

with probabilities

$$p_{1111111} = \tfrac{2}{7} \quad p_{0000000} = \tfrac{4}{7} \quad p_{1101111} = \tfrac{1}{42} \; p_{1111101} = \tfrac{1}{42}$$

$$p_{1111100} = \tfrac{1}{42} \; p_{1111011} = \tfrac{1}{42} \; p_{1100111} = \tfrac{1}{42} \; p_{0011111} = \tfrac{1}{42}$$

and all other $p_{a_1 \cdots a_7} = 0$.

We need to consider Flattenings of the boundary tensor $P = (p_{a_1 \cdots a_7})$ of the form

(1) $\mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}$
(2) $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6,\ell_7\}}$
(3) $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5,\ell_6,\ell_7\}}$
(4) $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}$
(5) $\mathrm{Flat}_{\{\ell_4,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_5\}}$
(6) $\mathrm{Flat}_{\{\ell_3,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_4,\ell_5\}}$

Note that we do not need to consider the flattenings $\mathrm{Flat}_{\{\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4,\ell_5\}}$ and $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6,\ell_6\}}$, as these are common to all the candidate trees and would not help discriminating between them.

All the flattenings above correspond to 8×16 matrices as in Fig. 1, where in each of the cases listed above the matrix indices $(abcdefg)$ correspond, respectively, to

(1) $(abcdefg) = (a_5a_6a_7a_1a_2a_3a_4)$
(2) $(abcdefg) = (a_1a_2a_3a_4a_5a_6a_7)$

$$\begin{pmatrix}
p_{0000000} & p_{0001000} & p_{0000100} & p_{0000010} & p_{0000001} & p_{0001100} & p_{0001010} & p_{0001001} & p_{0000110} & p_{0000101} & p_{0000011} & p_{0001110} & p_{0001101} & p_{0001011} & p_{0000111} & p_{0001111} \\
p_{1000000} & p_{1001000} & p_{1000100} & p_{1000010} & p_{1000001} & p_{1001100} & p_{1001010} & p_{1001001} & p_{1000110} & p_{1000101} & p_{1000011} & p_{1001110} & p_{1001101} & p_{1001011} & p_{1000111} & p_{1001111} \\
p_{0100000} & p_{0101000} & p_{0100100} & p_{0100010} & p_{0100001} & p_{0101100} & p_{0101010} & p_{0101001} & p_{0100110} & p_{0100101} & p_{0100011} & p_{0101110} & p_{0101101} & p_{0101011} & p_{0100111} & p_{0101111} \\
p_{0010000} & p_{0011000} & p_{0010100} & p_{0010010} & p_{0010001} & p_{0011100} & p_{0011010} & p_{0011001} & p_{0010110} & p_{0010101} & p_{0010011} & p_{0011110} & p_{0011101} & p_{0011011} & p_{0010111} & p_{0011111} \\
p_{0110000} & p_{0111000} & p_{0110100} & p_{0110010} & p_{0110001} & p_{0111100} & p_{0111010} & p_{0111001} & p_{0110110} & p_{0110101} & p_{0110011} & p_{0111110} & p_{0111101} & p_{0111011} & p_{0110111} & p_{0111111} \\
p_{1010000} & p_{1011000} & p_{1010100} & p_{1010010} & p_{1010001} & p_{1011100} & p_{1011010} & p_{1011001} & p_{1010110} & p_{1010101} & p_{1010011} & p_{1011110} & p_{1011101} & p_{1011011} & p_{1010111} & p_{1011111} \\
p_{1100000} & p_{1101000} & p_{1100100} & p_{1100010} & p_{1100001} & p_{1101100} & p_{1101010} & p_{1101001} & p_{1100110} & p_{1100101} & p_{1100011} & p_{1101110} & p_{1101101} & p_{1101011} & p_{1100111} & p_{1101111} \\
p_{1110000} & p_{1111000} & p_{1110100} & p_{1110010} & p_{1110001} & p_{1111100} & p_{1111010} & p_{1111001} & p_{1110110} & p_{1110101} & p_{1110011} & p_{1111110} & p_{1111101} & p_{1111011} & p_{1110111} & p_{1111111}
\end{pmatrix}$$

Fig. 1 Flattenings 8×16 matrices

(3) $(abcdefg) = (a_1a_2a_4a_3a_5a_6a_7)$
(4) $(abcdefg) = (a_1a_2a_5a_3a_4a_6a_7)$
(5) $(abcdefg) = (a_4a_6a_7a_1a_2a_3a_5)$
(6) $(abcdefg) = (a_3a_6a_7a_1a_2a_4a_5)$

The probability distributions corresponding to the permutations listed above are respectively given by

(1) $n_{1111101} = 1, n_{1011111} = 1, n_{1001111} = 1, n_{0111111} = 1, n_{1111100} = 1, n_{1110011} = 1$
(2) $n_{1101111} = 1, n_{1111101} = 1, n_{1111100} = 1, n_{1111011} = 1, n_{1100111} = 1, n_{0011111} = 1$
(3) $n_{1110111} = 1, n_{1111101} = 1, n_{1111100} = 1, n_{1111011} = 1, n_{1100111} = 1, n_{0011111} = 1$
(4) $n_{1110111} = 1, n_{1111101} = 1, n_{1111100} = 1, n_{1101111} = 1, n_{1110011} = 1, n_{0011111} = 1$
(5) $n_{1111101} = 1, n_{1011111} = 1, n_{1001111} = 1, n_{1111110} = 1, n_{0111101} = 1, n_{1110011} = 1$
(6) $n_{0111111} = 1, n_{1011111} = 1, n_{1001111} = 1, n_{1111110} = 1, n_{0111101} = 1, n_{1110011} = 1$

while all six cases have the common values $n_{1111111} = 12$ and $n_{0000000} = 24$.
The corresponding flattening matrices are given by

$$\mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}(P) = \begin{pmatrix}
\frac{4}{7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{42} & 0 & 0 & 0 & 0 & \frac{1}{42} & 0 & \frac{1}{42} & 0 & 0 & \frac{2}{7}
\end{pmatrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6,\ell_7\}}(P) = \begin{pmatrix}
\frac{4}{7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} & \frac{1}{42} \\
0 & 0 & 0 & 0 & 0 & \frac{1}{42} & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} & \frac{1}{42} & 0 & \frac{2}{7}
\end{pmatrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5,\ell_6,\ell_7\}}(P) = \begin{pmatrix} \frac{4}{7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{42} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} & \frac{1}{42} & \frac{1}{42} & \frac{2}{7} \end{pmatrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}(P) = \begin{pmatrix} \frac{4}{7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{42} & 0 & 0 & 0 & 0 & \frac{1}{42} & 0 & \frac{1}{42} & 0 & \frac{1}{42} & \frac{2}{7} \end{pmatrix}$$

$$\mathrm{Flat}_{\{\ell_4,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_5\}}(P) = \begin{pmatrix} \frac{4}{7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{42} & 0 & 0 & 0 & 0 & \frac{1}{42} & \frac{1}{42} & \frac{1}{42} & 0 & 0 & \frac{2}{7} \end{pmatrix}$$

$$\mathrm{Flat}_{\{\ell_3,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_4,\ell_5\}}(P) = \begin{pmatrix} \frac{4}{7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} & \frac{1}{42} & 0 & 0 & 0 & \frac{2}{7} \end{pmatrix}$$

The trees $T_5(G)$ and $T_6(G)$ have a slightly different structure, since in addition to placing in closest proximity the pairs $\{\ell_1, \ell_2\}$ and $\{\ell_6, \ell_7\}$ like all other trees they also identify pairs $\{\ell_4, \ell_5\}$ in the case of $T_5(G)$ and $\{\ell_3, \ell_5\}$ in the case of $T_6(G)$. Thus, while these two trees also have the flattenings $\text{Flat}_{\{\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4,\ell_5\}}$ and $\text{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6,\ell_6\}}$ common to all the other trees, they also have a flattening

$\text{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}$

common to both trees $T_5(G)$ and $T_6(G)$ and

$$\begin{array}{l} F_5 := \text{Flat}_{\{\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_3,\ell_6,\ell_7\}} \text{ for } T_5(G) \\ F_6 := \text{Flat}_{\{\ell_3,\ell_5\}\cup\{\ell_1,\ell_2,\ell_4,\ell_6,\ell_7\}} \text{ for } T_6(G). \end{array} \tag{3.1}$$

We have as corresponding matrices

$$\text{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}(P) = \begin{pmatrix} \frac{4}{7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{42} & 0 & \frac{1}{42} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{42} & \frac{2}{7} \end{pmatrix}$$

while the matrices (written in transpose form) for F_5 and F_6 are given in Appendix C.

3.10 Computation of the Phylogenetic Invariants

We compute the phylogenetic invariants using the ℓ^∞ and the ℓ^1 norm.

(1) The tree $T_1(G)$ with flattenings $M_1 = \text{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}$ and $M_2 = \text{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6,\ell_7\}}$ gives

$$\|\phi_{T_1}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}(M_1)} |\phi(P)| \,,\, \max_{\phi\in\mathcal{D}^{(3)}(M_2)} |\phi(P)|\} = \frac{4}{1029}$$

$$\|\phi_{T_1}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}(M_1)} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}(M_2)} |\phi(P)| = \frac{83}{8232}$$

(2) The tree $T_2(G)$ with flattenings $M_1 = \text{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}$ and $M_3 = \text{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5,\ell_6,\ell_7\}}$ gives

$$\|\phi_{T_2}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}(M_1)} |\phi(P)| \,,\, \max_{\phi\in\mathcal{D}^{(3)}(M_3)} |\phi(P)|\} = \frac{4}{1029}$$

$$\|\phi_{T_2}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}(M_1)} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}(M_3)} |\phi(P)| = \frac{233}{24696}$$

(3) The tree $T_3(G)$ with flattenings $M_4 = \text{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}$ and $M_5 = \text{Flat}_{\{\ell_4,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_5\}}$ gives

$$\|\phi_{T_3}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}(M_4)} |\phi(P)| \,,\, \max_{\phi\in\mathcal{D}^{(3)}(M_5)} |\phi(P)|\} = \frac{1}{3087}$$

$$\|\phi_{T_3}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}(M_4)} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}(M_5)} |\phi(P)| = \frac{16}{3087}$$

(4) The tree $T_4(G)$ with flattenings $M_4 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}$ and $M_6 = \mathrm{Flat}_{\{\ell_3,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_4,\ell_5\}}$ gives

$$\|\phi_{T_4}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}(M_4)} |\phi(P)| \,,\, \max_{\phi\in\mathcal{D}^{(3)}(M_6)} |\phi(P)|\} = \frac{4}{1029}$$

$$\|\phi_{T_4}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}(M_4)} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}(M_6)} |\phi(P)| = \frac{181}{18522}$$

(5) The tree $T_5(G)$ with flattenings F_5 of Appendix C and $M_7 = \mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}$ gives

$$\|\phi_{T_5}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}(F_5)} |\phi(P)| \,,\, \max_{\phi\in\mathcal{D}^{(3)}(M_7)} |\phi(P)|\} = \frac{4}{1029}$$

$$\|\phi_{T_5}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}(F_5)} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}(M_7)} |\phi(P)| = \frac{233}{24696}$$

(6) The tree $T_6(G)$ with flattenings F_6 of Appendix C and $M_7 = \mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}$ gives

$$\|\phi_{T_6}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}(F_6)} |\phi(P)| \,,\, \max_{\phi\in\mathcal{D}^{(3)}(M_7)} |\phi(P)|\} = \frac{4}{1029}$$

$$\|\phi_{T_6}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}(F_6)} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}(M_7)} |\phi(P)| = \frac{83}{8232}$$

In this case we see that both the ℓ^∞ and the ℓ^1 norm provide a good test that selects the historically correct tree $T_3(G)$. Note that the ℓ^∞ has the same value $4/1029$ on all the other candidates and the lower value $1/3087$ only for the correct tree $T_3(G)$.

3.11 Estimates of Euclidean Distance for the $\mathcal{S}_2(G)$ Germanic Languages

We obtain an evaluation of the candidate trees based on computing a lower bound for the Euclidean distance in terms of distances between the flattening matrices $\mathrm{Flat}_{e,T}(P)$ of the boundary distribution P and the determinantal varieties they are expected to lie on. As before, we use the notation with the explicit split of the leaves for the flattening matrices. More concretely, we have the following:

(1) The Euclidean distance estimate for the tree $T_1(G)$ is given by $\mathrm{dist}(P, V_{T_1}) \geq L_1$ with

$$L_1 = \max\{d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16)), d(\mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}(P), \mathcal{D}_2(8,16))\}$$

(2) The Euclidean distance estimate of $T_2(G)$ is given by $\mathrm{dist}(P, V_{T_2}) \geq L_2$ with

$$L_2 = \max\{d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16)), d(\mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}(P), \mathcal{D}_2(8,16))\}$$

(3) The Euclidean distance estimate of $T_3(G)$ is given by $\mathrm{dist}(P, V_{T_3}) \geq L_3$ with

$$L_3 = \max\{d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16)), d(\mathrm{Flat}_{\{\ell_4,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_5\}}(P), \mathcal{D}_2(8,16))\}$$

(4) The Euclidean distance estimate of $T_4(G)$ is given by $\mathrm{dist}(P, V_{T_4}) \geq L_4$ with

$$L_4 = \max\{d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16)), d(\mathrm{Flat}_{\{\ell_3,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_4,\ell_5\}}(P), \mathcal{D}_2(8,16))\}$$

(5) The Euclidean distance estimate of $T_5(G)$ is given by $\mathrm{dist}(P, V_{T_5}) \geq L_5$ with

$$L_5 = \max\{d(\mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16))^2, d(F_5(P), \mathcal{D}_2(4,32))^2\}$$

(6) The Euclidean distance estimate of $T_6(G)$ is given by $\mathrm{dist}(P, V_{T_6}) \geq L_6$ with

$$L_6 = \max\{d(\mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16))^2, d(F_6(P), \mathcal{D}_2(4,32))^2\}.$$

The singular value decomposition of the flattening matrices gives $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_8)$ with

$$\begin{aligned}
&\Sigma(\mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}(P))\\
&\quad\sim \mathrm{diag}(0.57143, 0.291548, 0.58333\times 10^{-2}, 0.12240\times 10^{-17},\\
&\quad 0.10572\times 10^{-34}, 0.16149\times 10^{-51}, 0.63652\times 10^{-68}, 0)\\
&\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6,\ell_7\}})(P))\\
&\quad\sim \mathrm{diag}(0.57143, 0.29059, 0.23973\times 10^{-1}, 0.33558\times 10^{-2}, 0.64145\times 10^{-19}, 0.60260\times 10^{-31}, 0, 0)\\
&\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5,\ell_6,\ell_7\}}(P))\\
&\quad\sim \mathrm{diag}(0.57143, 0.29061, 0.23809\times 10^{-1}, 0.33787\times 10^{-2}, 0, 0, 0, 0)\\
&\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}(P))\\
&\quad\sim \mathrm{diag}(0.57143, 0.29155, 0.54996\times 10^{-2}, 0, 0, 0, 0, 0)\\
&\Sigma(\mathrm{Flat}_{\{\ell_4,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_5\}}(P))\\
&\quad\sim \mathrm{diag}(0.57143, 0.29155, 0.54996\times 10^{-2}, 0, 0, 0, 0, 0)\\
&\Sigma(\mathrm{Flat}_{\{\ell_3,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_4,\ell_5\}}(P))\\
&\quad\sim \mathrm{diag}(0.57143, 0.29059, 0.23892\times 10^{-1}, 0.38881\times 10^{-2}, 0.12435\times 10^{-17}, 0.73417\times 10^{-19},\\
&\quad 0.32257\times 10^{-34}, 0).\\
&\Sigma(\mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}(P))\\
&\quad\sim \mathrm{diag}(0.57143, 0.29155, 0.58333\times 10^{-2}, 0.18608\times 10^{-17}, 0.32093\times 10^{-33}, 0, 0, 0)\\
&\Sigma(F_5(P)) = (0.57143, 0.29061, 0.23809\times 10^{-1}, 0.33787\times 10^{-2})\\
&\Sigma(F_6(P)) = (0.57143, 0.29060, 0.23973\times 10^{-1}, 0.33558\times 10^{-2})
\end{aligned}$$

By the Eckart-Young theorem we then have

$$\begin{aligned}
&d(\mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}(P), \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.34027\times 10^{-4}\\
&d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6,\ell_7\}})(P), \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.58597\times 10^{-3}\\
&d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.57831\times 10^{-3}\\
&d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.30245\times 10^{-4}\\
&d(\mathrm{Flat}_{\{\ell_4,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_5\}}(P), \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.30245\times 10^{-4}\\
&d(\mathrm{Flat}_{\{\ell_3,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_4,\ell_5\}}(P), \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.58595\times 10^{-3}\\
&d(\mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.34027\times 10^{-4}\\
&d(F_5(P), \mathcal{D}_2(4,32))^2 = \sigma_3^2 + \sigma_4^2 = 0.57831\times 10^{-3}\\
&d(F_6(P), \mathcal{D}_2(4,32))^2 = \sigma_3^2 + \sigma_4^2 = 0.58597\times 10^{-3}.
\end{aligned}$$

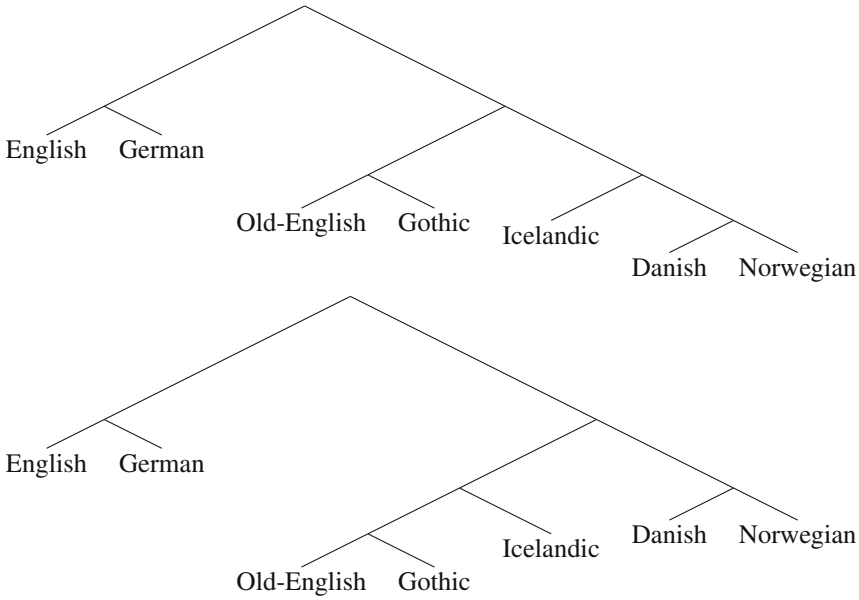
Thus, we obtain

$$L_1 = 0.58597\times 10^{-3},\quad L_2 = 0.57831\times 10^{-3},\quad L_3 = 0.30245\times 10^{-4},\quad L_4 = 0.58595\times 10^{-3},$$
$$L_5 = 0.57831\times 10^{-3},\quad L_6 = 0.58597\times 10^{-3}.$$

Thus, the computation of the phylogenetic invariants selects the tree $T_3(G)$ as the preferred candidate phylogenetic tree. The estimate of the Euclidean distance shows that the lower bounds obtained for the trees $T_1(G), T_2(G), T_4(G), T_5(G), T_6(G)$ are comparable and only $T_3(G)$ has a significantly smaller estimate. Thus, this criterion, even if it is only based on lower bounds, also suggests $T_3(G)$ as the most favorable candidate. The tree $T_3(G)$ is indeed the closest to what is regarded as the correct linguistic phylogenetic tree.

3.12 Comparison with SSWL Data

The DNA parsimony algorithm of PHYLIP produced the following two candidate phylogenetic trees for the set $\mathcal{S}_2(G)$ of Germanic languages based on the combination of the Longobardi data and the SSWL data.



In this case, the inclusion of the additional SSWL data resolves the ambiguity of the PHYLIP tree discussed in Sect. 3.9. In terms of our treatment of the positioning of the ancient languages, the two trees shown here should be regarded as corresponding to the possible trees in cases (3) and (4) discussed above in §3.9, for the first tree and cases (5) and (6) for the second one.

Thus, the set of possible binary trees we should consider for a comparison between the phylogenetic invariants evaluated on the Longobardi and on the SSWL data, consists of the trees $T_3(G)$ and $T_4(G)$ and $T_5(G)$ and $T_6(G)$ of the previous section. We will evaluate here the phylogenetic invariants and estimate the Euclidean distance function of these candidate trees (including for completeness also $T_1(G)$ and $T_2(G)$ of the previous section) using the boundary distribution based on the SSWL data.

3.13 Boundary Distribution for $\mathcal{S}_2(G)$ Based on SSWL Data

The Germanic languages in the set $\mathcal{S}_2(G)$ have a total of 68 SSWL variables that are completely mapped for all the seven languages in the set. This is significantly smaller than the 90 variables used for the set $\mathcal{S}_1(G)$. This does not depend on the languages being poorly mapped: the levels of accuracy are comparable with the previous set with Danish (76%), Norwegian (75%), German (75%), English (75%), Old English (75%) Icelandic (62%), Gothic (62%). However, the regions of the overall 115 SSWL variables that are mapped is less uniform across this set of languages creating a smaller overlap. The set of completely mapped SSWL variables for this set of languages is reported in Appendix B.

The occurrences of binary vectors at the leaves is given by

$$
\begin{array}{lll}
n_{0,0,0,0,0,0,0} = 26 & n_{1,1,1,1,1,1,1} = 16 & n_{0,0,1,1,0,0,1} = 2 \\
n_{0,0,1,0,0,0,0} = 3 & n_{1,1,0,1,0,0,0} = 1 & n_{0,0,1,1,1,1,0} = 1 \\
n_{0,0,1,1,1,0,0} = 1 & n_{0,0,1,0,1,0,0} = 1 & n_{1,1,0,1,0,1,1} = 2 \\
n_{1,0,1,1,1,0,0} = 1 & n_{1,1,1,1,1,0,1} = 1 & n_{1,1,1,1,1,0,0} = 1 \\
n_{1,1,1,1,0,1,1} = 3 & n_{1,1,0,1,1,0,1} = 1 & n_{0,0,0,0,1,0,0} = 1 \\
n_{1,1,0,0,1,1,1} = 1 & n_{0,0,0,0,0,1,0} = 1 & n_{0,0,0,1,0,0,0} = 2 \\
n_{0,0,0,0,0,0,1} = 1 & n_{0,0,1,1,0,0,0} = 1 & n_{1,1,0,1,1,1,1} = 1
\end{array}
$$

Thus, the boundary probability distribution for the SSWL data for these seven Germanic languages is given by

$$
\begin{array}{lll}
p_{0,0,0,0,0,0,0} = \frac{13}{34} & p_{1,1,1,1,1,1,1} = \frac{4}{17} & p_{0,0,1,1,0,0,1} = \frac{1}{34} \\
p_{0,0,1,0,0,0,0} = \frac{3}{68} & p_{1,1,0,1,0,0,0} = \frac{1}{68} & p_{0,0,1,1,1,1,0} = \frac{1}{68} \\
p_{0,0,1,1,1,0,0} = \frac{1}{68} & p_{0,0,1,0,1,0,0} = \frac{1}{68} & p_{1,1,0,1,0,1,1} = \frac{1}{34} \\
p_{1,0,1,1,1,0,0} = \frac{1}{68} & p_{1,1,1,1,1,0,1} = \frac{1}{68} & p_{1,1,1,1,1,0,0} = \frac{1}{68} \\
p_{1,1,1,1,0,1,1} = \frac{3}{68} & p_{1,1,0,1,1,0,1} = \frac{1}{68} & p_{0,0,0,0,1,0,0} = \frac{1}{68} \\
p_{1,1,0,0,1,1,1} = \frac{1}{68} & p_{0,0,0,0,0,1,0} = \frac{1}{68} & p_{0,0,0,1,0,0,0} = \frac{1}{34} \\
p_{0,0,0,0,0,0,1} = \frac{1}{68} & p_{0,0,1,1,0,0,0} = \frac{1}{68} & p_{1,1,0,1,1,1,1} = \frac{1}{68}
\end{array}
$$

The six flattening matrices corresponding to the different trees of the previous section are in this case of the following form.

$$
\mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}(P) = \begin{pmatrix}
\frac{13}{34} & 0 & 0 & \frac{3}{68} & \frac{1}{34} & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & \frac{1}{68} & 0 & 0 & 0 \\
\frac{1}{68} & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & \frac{1}{68} & 0 & \frac{1}{68} \\
\frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{34} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{34} & 0 & 0 & \frac{3}{68} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & \frac{1}{68} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & \frac{4}{17}
\end{pmatrix}
$$

$$
\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6,\ell_7\}}(P) = \begin{pmatrix}
\frac{13}{34} & \frac{1}{34} & \frac{1}{68} & \frac{1}{68} & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{3}{68} & \frac{1}{68} & \frac{1}{68} & 0 & 0 & \frac{1}{68} & 0 & \frac{1}{34} & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & \frac{1}{34} & \frac{1}{68} & \frac{1}{68} \\
0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & \frac{3}{68} & 0 & \frac{4}{17}
\end{pmatrix}
$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5,\ell_6,\ell_7\}}(P) = \begin{pmatrix} \frac{13}{34} & \frac{3}{68} & \frac{1}{68} & \frac{1}{68} & \frac{1}{68} & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{34} & \frac{1}{68} & 0 & 0 & 0 & \frac{1}{68} & 0 & \frac{1}{34} & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 \\ \frac{1}{68} & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & \frac{1}{68} & \frac{1}{34} & 0 & \frac{1}{68} & \frac{3}{68} & \frac{1}{68} & \frac{4}{17} \end{pmatrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}(P) = \begin{pmatrix} \frac{13}{34} & \frac{3}{68} & \frac{1}{34} & \frac{1}{68} & \frac{1}{68} & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{34} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{68} & \frac{1}{68} & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{34} & \frac{3}{68} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & \frac{1}{68} & \frac{1}{68} & 0 & \frac{1}{68} & 0 & \frac{1}{68} & \frac{4}{17} \end{pmatrix}$$

$$\mathrm{Flat}_{\{\ell_4,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_5\}}(P) = \begin{pmatrix} \frac{13}{34} & 0 & 0 & \frac{3}{68} & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{34} & 0 & 0 & \frac{1}{68} & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & \frac{1}{68} & 0 & \frac{1}{68} \\ \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{34} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & \frac{1}{68} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{34} & 0 & 0 & 0 & 0 & 0 & \frac{3}{68} & \frac{1}{68} & 0 & 0 & \frac{4}{17} \end{pmatrix}$$

$$\mathrm{Flat}_{\{\ell_3,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_4,\ell_5\}}(P) = \begin{pmatrix} \frac{13}{34} & 0 & 0 & \frac{1}{34} & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 \\ \frac{3}{68} & 0 & 0 & \frac{1}{68} & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & \frac{1}{68} & 0 & \frac{1}{68} \\ \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{34} & \frac{1}{68} & 0 & 0 & \frac{1}{68} \\ 0 & 0 & 0 & \frac{1}{34} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{68} & 0 & 0 & 0 & \frac{4}{17} \end{pmatrix}$$

3.13.1 The Trees T_5 and T_6

For the two remaining trees we have the flattening matrix

$$\mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}(P) = \begin{pmatrix} \frac{13}{34} & 0 & 0 & 0 & \frac{1}{68} & \frac{1}{68} & 0 & 0 \\ \frac{1}{34} & 0 & 0 & \frac{1}{68} & 0 & 0 & 0 & 0 \\ \frac{3}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{68} & 0 & 0 & 0 & 0 & \frac{1}{34} & 0 & 0 \\ \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{68} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{68} & \frac{1}{68} & 0 & \frac{1}{68} & \frac{1}{68} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{34} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{68} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & \frac{1}{68} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{68} & \frac{4}{17} \end{pmatrix}$$

and the matrices for the flattenings F_5 and F_6 given in the Appendix C.

3.14 Phylogenetic Invariants

We compute the phylogenetic invariants, using either the ℓ^∞ or the ℓ^1 norm. This case shows, as observed already in [8], that the ℓ^1 norm gives more reliable results than the ℓ^∞ norm.

- For the first tree $T_1(G)$ we consider all 3×3 minors of the flattenings

  $M_1 = \mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}(P)$ and $M_2 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6,\ell_7\}}(P)$

  and we obtain

  $$\|\phi_{T_1}(P)\|_{\ell^\infty} = \max_{\mathcal{D}^{(3)}(M_1)\cup\mathcal{D}^{(3)}(M_2)} |\phi(P)| = \frac{13}{4913}$$

  $$\|\phi_{T_1}(P)\|_{\ell^1} = \sum_{\mathcal{D}^{(3)}(M_1)\cup\mathcal{D}^{(3)}(M_2)} |\phi(P)| = \frac{8811}{157216}$$

- For the second tree $T_2(G)$ we consider all 3×3 minors of the flattenings

  $M_1 = \mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}(P)$ and $M_3 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5,\ell_6,\ell_7\}}(P)$

  and we obtain

  $$\|\phi_{T_2}(P)\|_{\ell^\infty} = \max_{\mathcal{D}^{(3)}(M_1)\cup\mathcal{D}^{(3)}(M_3)} |\phi(P)| = \frac{13}{4913}$$

  $$\|\phi_{T_2}(P)\|_{\ell^1} = \sum_{\mathcal{D}^{(3)}(M_1)\cup\mathcal{D}^{(3)}(M_3)} |\phi(P)| = \frac{7103}{157216}$$

- For the third tree $T_3(G)$ we consider all 3×3 minors of the flattenings

  $M_4 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}(P)$ and $M_5 = \mathrm{Flat}_{\{\ell_4,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_5\}}(P)$

  and we obtain

$$\|\phi_{T_3}(P)\|_{\ell^\infty} = \max_{\mathcal{D}^{(3)}(M_4)\cup\mathcal{D}^{(3)}(M_5)} |\phi(P)| = \frac{13}{4913}$$

$$\|\phi_{T_3}(P)\|_{\ell^1} = \sum_{\mathcal{D}^{(3)}(M_4)\cup\mathcal{D}^{(3)}(M_5)} |\phi(P)| = \frac{5439}{157216}$$

- For the fourth tree $T_4(G)$ we consider all 3×3 minors of the flattenings

  $M_4 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}(P)$ and $M_6 = \mathrm{Flat}_{\{\ell_3,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_4,\ell_5\}}(P)$

  and we obtain

$$\|\phi_{T_4}(P)\|_{\ell^\infty} = \max_{\mathcal{D}^{(3)}(M_4)\cup\mathcal{D}^{(3)}(M_6)} |\phi(P)| = \frac{13}{4913}$$

$$\|\phi_{T_4}(P)\|_{\ell^1} = \sum_{\mathcal{D}^{(3)}(M_4)\cup\mathcal{D}^{(3)}(M_6)} |\phi(P)| = \frac{5739}{157216}$$

- For the fifth tree $T_5(G)$ we consider all 3×3 minors of the flattenings

  $M_7 = \mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}(P)$ and F_5 (as in Appendix C)

  and we obtain

$$\|\phi_{T_5}(P)\|_{\ell^\infty} = \max_{\mathcal{D}^{(3)}(M_7)\cup\mathcal{D}^{(3)}(F_5)} |\phi(P)| = \frac{13}{4913}$$

$$\|\phi_{T_5}(P)\|_{\ell^1} = \sum_{\mathcal{D}^{(3)}(M_7)\cup\mathcal{D}^{(3)}(F_5)} |\phi(P)| = \frac{25}{578}$$

- For the sixth tree $T_6(G)$ we consider all 3×3 minors of the flattenings

  $M_7 = \mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}(P)$ and F_6 (as in Appendix C)

  and we obtain

$$\|\phi_{T_6}(P)\|_{\ell^\infty} = \max_{\mathcal{D}^{(3)}(M_7)\cup\mathcal{D}^{(3)}(F_6)} |\phi(P)| = \frac{207}{78608}$$

$$\|\phi_{T_6}(P)\|_{\ell^1} = \sum_{\mathcal{D}^{(3)}(M_7)\cup\mathcal{D}^{(3)}(F_6)} |\phi(P)| = \frac{11795}{314432}$$

When we evaluate the minimum among these candidate trees we see that using the ℓ^∞ norm in this case would incorrectly select the tree $T_6(G)$ as the best candidate, while using the ℓ^1 norm correctly selects $T_3(G)$

$$\min_T \|\phi_T(P)\|_{\ell^\infty} = \frac{207}{78608} = \|\phi_{T_6}(P)\|_{\ell^\infty}$$

$$\min_T \|\phi_T(P)\|_{\ell^1} = \frac{5439}{157216} = \|\phi_{T_3}(P)\|_{\ell^1}.$$

The ℓ^∞ norm also does not distinguish at all between the trees $T_1(G), \ldots, T_5(G)$.

3.15 Euclidean Distance Function Estimates

The Euclidean distance lower bound estimate can be obtained as in §3.11 by replacing the boundary probability based on the Longobardi data with the one based on SSWL data. We obtain the following.

The singular value decompositions $\Sigma = \mathrm{diag}(\sigma_k)$ are now of the form

$$\Sigma(M_1) = (0.38754, 0.24162, 0.36255 \times 10^{-1}, 0.29457 \times 10^{-1}, 0.17913 \times 10^{-1}, 0.18822 \times 10^{-2}, 0.44554 \times 10^{-3}, 0.81454 \times 10^{-18})$$

$$\Sigma(M_2) = (0.38705, 0.24121, 0.40755 \times 10^{-1}, 0.35206 \times 10^{-1}, 0.13458 \times 10^{-1}, 0.25922 \times 10^{-17}, 0.30537 \times 10^{-18}, 0.12727 \times 10^{-32})$$

$$\Sigma(M_3) = (0.38779, 0.24265, 0.37646 \times 10^{-1}, 0.14679 \times 10^{-1}, 0.13520 \times 10^{-1}, 0.72298 \times 10^{-17}, 0.10019 \times 10^{-18}, 0.15015 \times 10^{-30})$$

$$\Sigma(M_4) = (0.38833, 0.23760, 0.54943 \times 10^{-1}, 0.25989 \times 10^{-1}, 0.11091 \times 10^{-1}, 0.37355 \times 10^{-17}, 0.11876 \times 10^{-18}, 0.41814 \times 10^{-32})$$

$$\Sigma(M_5) = (0.38730, 0.24267, 0.35401 \times 10^{-1}, 0.25107 \times 10^{-1}, 0.13409 \times 10^{-1}, 0.10671 \times 10^{-1}, 0.83305 \times 10^{-3}, 0.63417 \times 10^{-18})$$

$$\Sigma(M_6) = (0.38735, 0.24147, 0.34918 \times 10^{-1}, 0.29212 \times 10^{-1}, 0.23098 \times 10^{-1}, 0.10765 \times 10^{-1}, 0.17668 \times 10^{-2}, 0.31311 \times 10^{-3})$$

$$\Sigma(M_7) = (0.38775, 0.24257, 0.29048 \times 10^{-1}, 0.26515 \times 10^{-1}, 0.14181 \times 10^{-1}, 0.11708 \times 10^{-1}, 0.13047 \times 10^{-2}, 0.60234 \times 10^{-18})$$

$$\Sigma(F_5) = (0.38710, 0.24296, 0.44347 \times 10^{-1}, 0.15179 \times 10^{-1})$$

$$\Sigma(F_6) = (0.39170, 0.23723, 0.30854 \times 10^{-1}, 0.20237 \times 10^{-1})$$

One obtains from these the Euclidean distances

$$d(M_1, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.25068 \times 10^{-2}$$
$$d(M_2, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.30816 \times 10^{-2}$$
$$d(M_3, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.18155 \times 10^{-2}$$
$$d(M_4, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.38172 \times 10^{-2}$$
$$d(M_5, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.21780 \times 10^{-2}$$
$$d(M_6, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.27252 \times 10^{-2}$$
$$d(M_7, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.18867 \times 10^{-2}$$
$$d(F_5, \mathcal{D}_2(4,32))^2 = \sigma_3^2 + \sigma_4^2 = 0.21971 \times 10^{-2}$$
$$d(F_6, \mathcal{D}_2(4,32))^2 = \sigma_3^2 + \sigma_4^2 = 0.13615 \times 10^{-2}.$$

From these distances, computed using the Eckart–Young theorem, one derives then estimates for the Euclidean distance of the form $\mathrm{dist}(P, V_{T_i}) \geq L_i$ where the L_i are computed as maxima of the distances in the list above that occur in the case of the tree T_i, in the same way as shown in Sect. 3.11.

We find that, in the case of the SSWL data for these Germanic languages, the lower bound on the Euclidean distance gives a less reliable answer. While it correctly excludes the candidates $T_1(G), T_2(G), T_4(G), T_5(G)$, it assigns the lowest value to the tree $T_6(G)$ rather than to the correct tree $T_3(G)$ selected by the phylogenetic invariants (computed with the ℓ^1-norm). Thus, we see here an example where the lower bound is an unreliable predictor of the actual Euclidean distance. This example confirms the expectation that Longobardi's LanGeLin data behave better for phylogenetic reconstruction than the SSWL data.

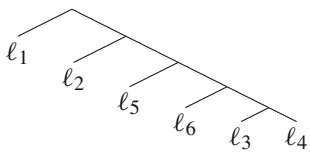
A possible explanation for this phenomenon lies in the fact that, although the list of SSWL variables for this set of languages is longer than the list of variables in the Longobardi data, there is a high degree of dependency between the SSWL data. This was also observed in [38] where the dependencies between SSWL variables were studied using Kanerva networks. Thus, the actual number of independent variables that contribute to the boundary distribution may be smaller in the use of the SSWL data. The fact that the languages in the set $\mathcal{S}_2(G)$ have a smaller overlap in the regions of the SSWL variables that are uniformly mapped for all languages, compared to those in the set $\mathcal{S}_1(G)$ further explains why the ℓ^∞-phylogenetic invariants and the Euclidean distance evaluated on the

boundary distribution of SSWL data correctly identify the best tree in the $\mathcal{S}_1(G)$ case but not in the $\mathcal{S}_2(G)$ case and the ℓ^1-phylogenetic invariant identifies the correct tree in the case of $\mathcal{S}_2(G)$ only by a small margin. We will return to discuss this point in §8 below.

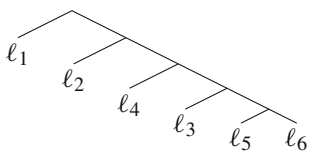
4 Phylogenetic Algebraic Varieties of the Romance Languages

We consider here the case of the Romance subfamily of the Indo-European language family. In particular, we focus of the relative position of the languages $\ell_1 =$ Latin, $\ell_2 =$ Romanian, $\ell_3 =$ French, $\ell_4 =$ Italian, $\ell_5 =$ Spanish, and $\ell_6 =$ Portuguese. We use the combined data of the SSWL and the Longobardi databases for this phylogenetic analysis, where we retain only those features of the SSWL database that are completely mapped for all of these languages.

When run on this set of syntactic data, the PHYLIP phylogenetic program produces a unique most parsimonious tree candidate, which is given by the tree T_1



with the additional linguistic information that ℓ_1 (Latin) should be considered as the root vertex, since the tree produced by PHYLIP is unrooted. There is clearly a problem with this tree, since the topology one expects based on historical linguistics is instead given by the tree T_2



4.1 Flattening Matrices of the PHYLIP Tree

There are three flattening matrices associated to the tree T_1, given by the three possible splits $e_1 = \{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5, \ell_6\}$, $e_2 = \{\ell_1, \ell_2, \ell_5\} \cup \{\ell_3, \ell_4, \ell_6\}$ and $e_3 = \{\ell_1, \ell_2, \ell_5, \ell_6\} \cup \{\ell_3, \ell_4\}$. With the boundary probability distribution given by the combined SSWL and Longobardi data, these are given by

$$\mathrm{Flat}_{e_1,T_1} = \begin{pmatrix} 0.2 & 0.0121 & 0.0606 & 0.0121 \\ 0 & 0 & 0 & 0.0061 \\ 0 & 0 & 0.0061 & 0 \\ 0 & 0 & 0.0061 & 0 \\ 0 & 0 & 0 & 0.0061 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0182 \\ 0.0242 & 0 & 0.0182 & 0 \\ 0 & 0.0061 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0061 & 0 \\ 0.0061 & 0 & 0 & 0.0061 \\ 0 & 0 & 0 & 0 \\ 0.0061 & 0 & 0 & 0.0061 \\ 0.0364 & 0.1091 & 0.0364 & 0.4121 \end{pmatrix}$$

$$\text{Flat}_{e_2,T_1} = \begin{pmatrix} 0.2 & 0 & 0.0121 & 0 & 0.0606 & 0 & 0.0121 & 0.0061 \\ 0 & 0 & 0 & 0 & 0.0061 & 0.0061 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0061 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0182 \\ 0.0242 & 0 & 0 & 0.0061 & 0.0182 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0061 & 0 & 0 \\ 0.0061 & 0 & 0 & 0 & 0 & 0 & 0.0061 & 0 \\ 0.0061 & 0.0364 & 0 & 0.1091 & 0 & 0.0364 & 0.0061 & 0.4121 \end{pmatrix}$$

while the third flattening Flat_{e_3,T_1} is given by

$$\begin{pmatrix} 0.2 & 0 & 0.0121 & 0 & 0 & 0 & 0 & 0 & 0.0606 & 0 & 0.0121 & 0.0061 & 0.0061 & 0.0061 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0061 & 0 & 0 & 0 & 0 & 0.0182 \\ 0.0242 & 0 & 0 & 0.0061 & 0 & 0 & 0 & 0 & 0.0182 & 0 & 0 & 0 & 0 & 0.0061 & 0 & 0 \\ 0.0061 & 0 & 0 & 0 & 0.0061 & 0.0364 & 0 & 0.1091 & 0 & 0 & 0.0061 & 0 & 0 & 0.0364 & 0.0061 & 0.4121 \end{pmatrix}$$

4.2 Flattening Matrices of the Historically Correct Tree

When we consider the linguistically correct tree T_2, instead of the tree T_1 computed by PHYLIP, using the same syntactic data for the boundary distribution, we find the flattening matrices which correspond to the splittings $e'_1 = \{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5, \ell_6\}$, $e'_2 = \{\ell_1, \ell_2, \ell_4\} \cup \{\ell_3, \ell_5, \ell_6\}$ and $e'_3 = \{\ell_1, \ell_2, \ell_3, \ell_4\} \cup \{\ell_5, \ell_6\}$.

$$\text{Flat}_{e'_1,T_2} = \begin{pmatrix} 0.2 & 0 & 0 & 0 \\ 0.0121 & 0 & 0 & 0 \\ 0.0606 & 0 & 0.0061 & 0.0061 \\ 0.0121 & 0.0061 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.0061 & 0 & 0 & 0.0182 \\ 0.0242 & 0 & 0 & 0 \\ 0 & 0.0061 & 0 & 0 \\ 0.0182 & 0 & 0 & 0.0061 \\ 0 & 0 & 0 & 0 \\ 0.0061 & 0 & 0.0061 & 0.0364 \\ 0 & 0 & 0 & 0.1091 \\ 0 & 0 & 0 & 0.0364 \\ 0.0061 & 0 & 0.0061 & 0.4121 \end{pmatrix}$$

$$\text{Flat}_{e'_2,T_2} = \begin{pmatrix} 0.2 & 0 & 0 & 0 & 0.0242 & 0 & 0 & 0 \\ 0.0121 & 0 & 0 & 0 & 0 & 0.0061 & 0 & 0 \\ 0.0606 & 0 & 0.0061 & 0.0061 & 0.0182 & 0 & 0 & 0.0061 \\ 0.0121 & 0.0061 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0061 & 0 & 0.0061 & 0.0364 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1091 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0364 \\ 0.0061 & 0 & 0 & 0.0182 & 0.0061 & 0 & 0.0061 & 0.4121 \end{pmatrix}$$

and with the third flattening matrix $\text{Flat}_{e'_3,T_2}$ given by

$$\begin{pmatrix} 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0242 & 0 & 0 & 0 & 0.0061 & 0 & 0.0061 & 0.0364 \\ 0.0121 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0061 & 0 & 0 & 0 & 0 & 0 & 0.1091 \\ 0.0606 & 0 & 0.0061 & 0.0061 & 0 & 0 & 0 & 0 & 0.0182 & 0 & 0 & 0.0061 & 0 & 0 & 0 & 0.0364 \\ 0.0121 & 0.0061 & 0 & 0 & 0.0061 & 0 & 0 & 0.0182 & 0 & 0 & 0 & 0 & 0.0061 & 0 & 0.0061 & 0.4121 \end{pmatrix}$$

4.3 Phylogenetic Invariants

We compare the phylogenetic invariants of these two trees computed with respect to the ℓ^∞ and the ℓ^1 norm.

(1) from the PHYLIP tree T_1 we obtain:

$$\|\Phi_{T_1}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}_{e_1,T_1}} |\phi(P)|, \max_{\phi\in\mathcal{D}^{(3)}_{e_2,T_1}} |\phi(P)|, \max_{\phi\in\mathcal{D}^{(3)}_{e_3,T_1}} |\phi(P)|\} = 0.89579\times 10^{-3}$$

$$\|\Phi_{T_1}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}_{e_1,T_1}} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}_{e_2,T_1}} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}_{e_3,T_1}} |\phi(P)| = 0.24790\times 10^{-1}$$

(2) for the historically correct tree T_2 we find:

$$\|\Phi_{T_2}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}_{e'_1,T_2}} |\phi(P)|, \max_{\phi\in\mathcal{D}^{(3)}_{e'_2,T_2}} |\phi(P)|, \max_{\phi\in\mathcal{D}^{(3)}_{e'_3,T_2}} |\phi(P)|\} = 0.89579\times 10^{-3}$$

$$\|\Phi_{T_2}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}_{e'_1,T_2}} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}_{e'_2,T_2}} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}_{e'_3,T_2}} |\phi(P)| = 0.22681\times 10^{-1}$$

Once again we see that the ℓ^1 norm reliably distinguishes the historically correct tree T_2 over the incorrect PHYLIP candidate, while the ℓ^∞ norm gives the same result for both candidate trees and does not help distinguishing them.

4.4 Estimate of the Euclidean Distance

We also compute a lower bound estimate on the Euclidean distance. In the case of the first tree T_1 The Euclidean distances of the flattening matrices from the respective determinantal varieties are given by

$$D_{1,1} = \mathrm{dist}(\mathrm{Flat}_{e_1,T_1}, \mathcal{D}_2(4,16)), \quad D_{1,2} = \mathrm{dist}(\mathrm{Flat}_{e_2,T_1}, \mathcal{D}_2(8,8)), \quad D_{1,3} = \mathrm{dist}(\mathrm{Flat}_{e_3,T_1}, \mathcal{D}_2(16,4)).$$

The singular values of the flattening matrices are given, respectively, by

$$\Sigma(\mathrm{Flat}_{e_1,T_1}) = (0.4320, 0.2075, 0.14766\times 10^{-1}, 0.8211\times 10^{-2})$$

while the singular values of Flat_{e_2,T_1} are given by

$$(0.4299, 0.2115, 0.1390\times 10^{-1}, 0.8586\times 10^{-2}, 0.7806\times 10^{-2}, 0.4896\times 10^{-2}, 0.8464\times 10^{-3}, 0.1867\times 10^{-3})$$

and

$$\Sigma(\mathrm{Flat}_{e_3,T_1}) = (0.4299, 0.2118, 0.1332\times 10^{-1}, 0.7593\times 10^{-2}).$$

Thus, the Euclidean distances are given, respectively, by

$$D^2_{1,1} = 0.2854\times 10^{-3}$$
$$D^2_{1,2} = 0.3525\times 10^{-3}$$
$$D^2_{1,3} = 0.2351\times 10^{-3}$$

For the second tree T_2 the Euclidean distances of the flattening matrices to the corresponding determinantal varieties are given by

$$D^2_{2,1} = 0.1390\times 10^{-3},$$

which is computed using the singular values

$$\Sigma(\mathrm{Flat}_{e_1,T_2}) = (0.4300, 0.2119, 0.8567\times 10^{-2}, 0.8102\times 10^{-2}),$$
$$D^2_{2,2} = 0.3390\times 10^{-3}$$

computed using the singular values $\Sigma(\text{Flat}_{e_2,T_2})$ given by

$(0.4299, 0.2115, 0.14218 \times 10^{-1}, 0.6889 \times 10^{-2}, 0.6061 \times 10^{-2}, 0.6007 \times 10^{-2}, 0.4070 \times 10^{-2}, 0.7823 \times 10^{-19})$

and

$D^2_{2,3} = 0.2854 \times 10^{-3}$

with singular values

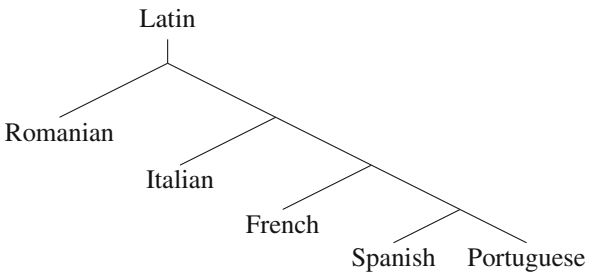
$\Sigma(\text{Flat}_{e_3,T_2}) = (0.4320, 0.2075, 0.1477 \times 10^{-1}, 0.8211 \times 10^{-2})$.

Thus if we compare the two models T_1 and T_2 using the maximum between the distances as a lower bound for the Euclidean distance to the phylogenetic variety we find

$L_1 = \max\{D^2_{1,1}, D^2_{1,2}, D^2_{1,3}\} = 0.3525 \times 10^{-3}$

$L_2 = \max\{D^2_{2,1}, D^2_{2,2}, D^2_{2,3}\} = 0.3390 \times 10^{-3}$,

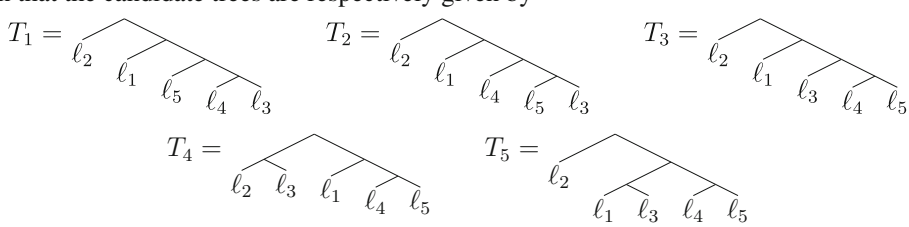
hence $L_2 < L_1$, which also favors the historically correct tree T_2:



5 Phylogenetic Algebraic Varieties of the Slavic Languages

We then consider a set of Slavic languages: $\ell_1 =$ Russian, $\ell_2 =$ Polish, $\ell_3 =$ Bulgarian, $\ell_4 =$ Serb-Croatian, $\ell_5 =$ Slovenian, for which we again use a combination of SSWL and Longobardi data. The PHYLIP most parsimonious trees algorithm produces in this case five candidate trees when run on this combination of syntactic data. We use additional linguistic information on where the root vertex should be placed, separating the West-Slavic branch where Polish resides from the part of the tree that contains both the East-Slavic branch and the South-Slavic branch.

We see then that the candidate trees are respectively given by



(1) The first tree T_1 incorrectly places Bulgarian in closer proximity to Serb-Croatian than Slovenian.
(2) The second tree T_2 has a similar misplacement, with Bulgarian appearing to be in greater proximity to Slovenian than Serb-Croatian.
(3) The third tree T_3 correctly places Slovenian and Serb-Croatian in closest proximity, and it also correctly places Bulgarian in the same South-Slavic subbranch with the pair of Slovenian and Serb-Croatian, so it corresponds to the correct tree topology that matches what is known from historical linguistics.
(4) The fourth tree T_4 misplaces Bulgarian in the West-Slavic branch with Polish instead of placing it in the South-Slavic branch.
(5) The fifth tree T_5 misplaces Bulgarian in the East-Slavic branch with Russian instead of placing it in the South-Slavic branch.

5.1 Flattening Matrices

We write here the flattening matrices using either the edge and tree subscript of the split notation as in Sect. 3, according to how it is more convenient: the following list makes it clear how these two notations match. The splits for the trees above are given by

$$\begin{aligned}
&T_1 : e_1 = \{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5\},\ e_2 = \{\ell_1, \ell_2, \ell_5\} \cup \{\ell_3, \ell_4\} \\
&T_2 : e_1 = \{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5\},\ e_2 = \{\ell_1, \ell_2, \ell_4\} \cup \{\ell_3, \ell_5\} \\
&T_3 : e_1 = \{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5\},\ e_2 = \{\ell_1, \ell_2, \ell_3\} \cup \{\ell_4, \ell_5\} \\
&T_4 : e_1 = \{\ell_2, \ell_3\} \cup \{\ell_1, \ell_4, \ell_5\},\ e_2 = \{\ell_1, \ell_2, \ell_3\} \cup \{\ell_4, \ell_5\} \\
&T_5 : e_1 = \{\ell_1, \ell_3\} \cup \{\ell_2, \ell_4, \ell_5\},\ e_2 = \{\ell_1, \ell_2, \ell_3\} \cup \{\ell_4, \ell_5\}.
\end{aligned} \tag{5.1}$$

The flattening matrices for these trees are given by the following

(1) For the tree T_1 the flattening matrices are

$$\mathrm{Flat}_{e_1,T_1} = \mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5\}} = \begin{pmatrix} 0.5122 & 0.0 & 0.0122 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0122 & 0.0 & 0.0 & 0.0610 \\ 0.0854 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0122 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.3049 \end{pmatrix}$$

$$\mathrm{Flat}_{e_2,T_1} = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4\}} = \begin{pmatrix} 0.5122 & 0.0 & 0.0 & 0.0 & 0.0122 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0122 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0610 \\ 0.0854 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0122 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3049 \end{pmatrix}$$

(2) For the tree T_2 the flattening matrices are $\mathrm{Flat}_{e_1,T_2} = \mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5\}}$ as above and

$$\mathrm{Flat}_{e_2,T_2} = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5\}} = \begin{pmatrix} 0.5122 & 0.0 & 0.0854 & 0.0 \\ 0.0 & 0.0122 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0122 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0122 \\ 0.0 & 0.0610 & 0.0 & 0.3049 \end{pmatrix}$$

(3) For the tree T_3 the flattening matrices are $\mathrm{Flat}_{e_1,T_3} = \mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5\}}$ as above and

$$\mathrm{Flat}_{e_2,T_3} = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5\}} = \begin{pmatrix} 0.5122 & 0.0 & 0.0122 & 0.0 & 0.0854 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0122 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0122 & 0.0 & 0.0 & 0.0610 & 0.0 & 0.0 & 0.0 & 0.3049 \end{pmatrix}$$

(4) For the tree T_4 the flattening matrices are $\mathrm{Flat}_{e_2,T_4} = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5\}}$ as above and

$$\mathrm{Flat}_{e_1,T_4} = \mathrm{Flat}_{\{\ell_2,\ell_3\}\cup\{\ell_1,\ell_4,\ell_5\}} = \begin{pmatrix} 0.5122 & 0.0122 & 0.0854 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0122 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0122 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0610 & 0.0 & 0.3049 \end{pmatrix}$$

(5) For the tree T_5 the flattening matrices are $\mathrm{Flat}_{e_2,T_5} = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5\}}$ as above and

$$\mathrm{Flat}_{e_1,T_5} = \mathrm{Flat}_{\{\ell_1,\ell_3\}\cup\{\ell_2,\ell_4,\ell_5\}} = \begin{pmatrix} 0.5122 & 0.0 & 0.0854 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0122 & 0.0 & 0.0 & 0.0 \\ 0.0122 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0122 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0610 & 0.0 & 0.3049 \end{pmatrix}$$

5.2 Phylogenetic Invariants

When evaluating the phylogenetic invariant for the boundary probability distribution given by the combination of the SSWL and Longobardi data we have the following result

(1) For the tree T_1:

$$\|\Phi_{T_1}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}_{e_1,T_1}} |\phi(P)|, \max_{\phi\in\mathcal{D}^{(3)}_{e_2,T_1}} |\phi(P)|\} = 0.19043\times 10^{-2}$$

$$\|\Phi_{T_1}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}_{e_1,T_1}} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}_{e_2,T_1}} |\phi(P)| = 0.31794\times 10^{-2}$$

(2) For the tree T_2:

$$\|\Phi_{T_2}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}_{e_1,T_2}} |\phi(P)|, \max_{\phi\in\mathcal{D}^{(3)}_{e_2,T_2}} |\phi(P)|\} = 0.19043\times 10^{-2}$$

$$\|\Phi_{T_2}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}_{e_1,T_2}} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}_{e_2,T_2}} |\phi(P)| = 0.36582\times 10^{-2}$$

(3) For the tree T_3:

$$\|\Phi_{T_3}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}_{e_1,T_3}} |\phi(P)|, \max_{\phi\in\mathcal{D}^{(3)}_{e_2,T_3}} |\phi(P)|\} = 0.38087\times 10^{-3}$$

$$\|\Phi_{T_3}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}_{e_1,T_3}} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}_{e_2,T_3}} |\phi(P)| = 0.90864\times 10^{-3}$$

(4) For the tree T_4:

$$\|\Phi_{T_4}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}_{e_1,T_4}} |\phi(P)|, \max_{\phi\in\mathcal{D}^{(3)}_{e_2,T_4}} |\phi(P)|\} = 0.38087\times 10^{-3}$$

$$\|\Phi_{T_4}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}_{e_1,T_4}} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}_{e_2,T_4}} |\phi(P)| = 0.13621\times 10^{-2}$$

(5) For the tree T_5:

$$\|\Phi_{T_5}(P)\|_{\ell^\infty} = \max\{\max_{\phi\in\mathcal{D}^{(3)}_{e_1,T_5}} |\phi(P)|, \max_{\phi\in\mathcal{D}^{(3)}_{e_2,T_5}} |\phi(P)|\} = 0.38087\times 10^{-3}$$

$$\|\Phi_{T_5}(P)\|_{\ell^1} = \sum_{\phi\in\mathcal{D}^{(3)}_{e_1,T_5}} |\phi(P)| + \sum_{\phi\in\mathcal{D}^{(3)}_{e_2,T_5}} |\phi(P)| = 0.17175\times 10^{-2}$$

For this set of languages we see again, as observed in [8], that the ℓ^1 norm is a better test than the ℓ^∞ norm for the evaluation of the phylogenetic invariants. While the ℓ^∞ norm does not distinguish between the trees T_3, T_4, T_5, the ℓ^1 norm correctly singles out T_3 as the preferred candidate.

5.3 Estimates of Euclidean Distance

The matrix $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5\}}$ has singular values

$$\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5\}}) = (0.5195, 0.3111, 0.2023\times 10^{-2}, 0.2577\times 10^{-17}, 0, 0, 0, 0).$$

The matrix $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5\}}$ has singular values

$$\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5\}}) = (0.5196, 0.3110, 0.2391\times 10^{-2}, 0).$$

The remaining matrices have

$$\Sigma(\mathrm{Flat}_{e_2,T_1}) = (0.5194, 0.3112, 0.1196\times 10^{-1}, 0.2003\times 10^{-2}),$$
$$\Sigma(\mathrm{Flat}_{e_2,T_2}) = (0.5194, 0.3112, 0.1220\times 10^{-1}, 0.2004\times 10^{-2}, 0, 0, 0, 0),$$
$$\Sigma(\mathrm{Flat}_{e_1,T_4}) = (0.5195, 0.3111, 0.2438\times 10^{-2}, 0.1964\times 10^{-2}, 0, 0, 0, 0),$$
$$\Sigma(\mathrm{Flat}_{e_1,T_5}) = (0.5195, 0.3111, 0.2834\times 10^{-2}, 0.2390\times 10^{-2}, 0, 0, 0, 0).$$

The computation of the Euclidean distances then gives

(1) For the tree T_1

$$\mathrm{dist}(\mathrm{Flat}_{e_1,T_1}, \mathcal{D}_2(4,8))^2 = \sigma_3^2 + \cdots \sigma_8^2 = 0.4094\times 10^{-5}$$
$$\mathrm{dist}(\mathrm{Flat}_{e_2,T_1}, \mathcal{D}_2(8,4))^2 = \sigma_3^2 + \sigma_4^2 = 0.1470\times 10^{-3}$$

(2) For the tree T_2

$$\mathrm{dist}(\mathrm{Flat}_{e_1,T_2}, \mathcal{D}_2(4,8))^2 = \sigma_3^2 + \cdots \sigma_8^2 = 0.4094\times 10^{-5}$$
$$\mathrm{dist}(\mathrm{Flat}_{e_2,T_2}, \mathcal{D}_2(4,8))^2 = \sigma_3^2 + \cdots \sigma_8^2 = 0.1527\times 10^{-3}$$

(3) For the tree T_3

$$\mathrm{dist}(\mathrm{Flat}_{e_1,T_3}, \mathcal{D}_2(4,8))^2 = \sigma_3^2 + \cdots \sigma_8^2 = 0.4094\times 10^{-5}$$
$$\mathrm{dist}(\mathrm{Flat}_{e_2,T_3}, \mathcal{D}_2(4,8))^2 = \sigma_3^2 + \cdots \sigma_8^2 = 0.5718\times 10^{-5}$$

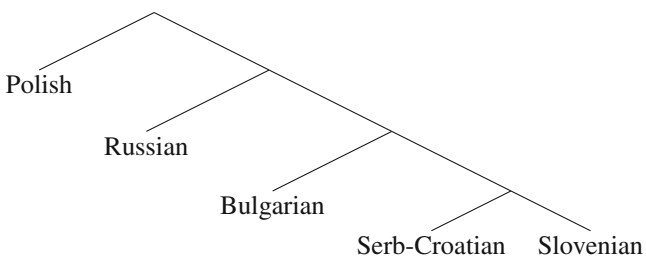
(4) For the tree T_4

$$\mathrm{dist}(\mathrm{Flat}_{e_1,T_4}, \mathcal{D}_2(4,8))^2 = \sigma_3^2 + \cdots \sigma_8^2 = 0.9803\times 10^{-5}$$
$$\mathrm{dist}(\mathrm{Flat}_{e_2,T_4}, \mathcal{D}_2(4,8))^2 = \sigma_3^2 + \cdots \sigma_8^2 = 0.5718\times 10^{-5}$$

(5) For the tree T_5

$$\mathrm{dist}(\mathrm{Flat}_{e_1,T_5}, \mathcal{D}_2(4,8))^2 = \sigma_3^2 + \cdots \sigma_8^2 = 0.1374\times 10^{-4}$$
$$\mathrm{dist}(\mathrm{Flat}_{e_2,T_5}, \mathcal{D}_2(4,8))^2 = \sigma_3^2 + \cdots \sigma_8^2 = 0.5718\times 10^{-5}$$

The lower bounds on the Euclidean distance function obtained above indicate as preferred candidate the tree T_3, which is the correct linguistic tree:

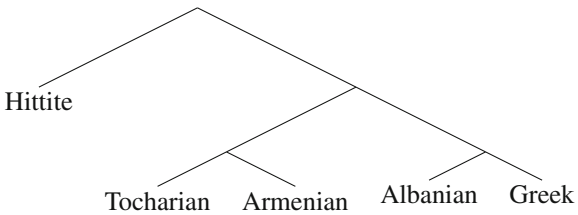


6 Phylogenetic Algebraic Varieties of the Early Indo-European Tree

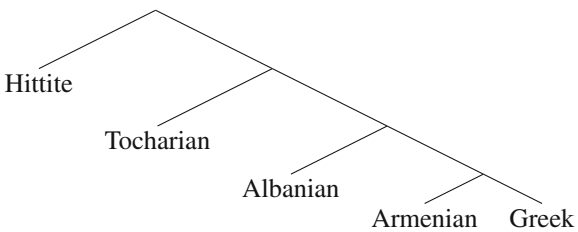
We now discuss the last phylogenetic problem listed in the Introduction, namely the early branchings of the Indo-European tree involving the set of languages Hittite, Tocharian, Albanian, Armenian, and Greek. We analyze here the difference between the trees of [6] and [43], when seen from the point of view of Phylogenetic Algebraic Geometry.

6.1 Trees and Phylogenetic Invariants

Once we restrict our attention to the five languages listed above, the trees of [6] and [43] that we wish to compare result in the smaller five-leaf trees

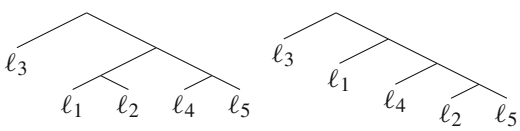


for the case computed by [6], and the tree



for the case computed by [43].

Forgetting momentarily the position of the root vertex (which is in both trees adjacent to the Anatolian branch), we are comparing two trees of the form



where we have $\ell_1 =$ Tocharian, $\ell_2 =$ Armenian, $\ell_3 =$ Hittite, $\ell_4 =$ Albanian, $\ell_5 =$ Greek. The splits correspond to

$$T_1 : e_1 = \{\ell_1, \ell_2\} \cup \{\ell_3, \ell_4, \ell_5\}\ e_2 = \{\ell_1, \ell_2, \ell_3\} \cup \{\ell_4, \ell_5\}$$
$$T_2 : e_1 = \{\ell_1, \ell_3\} \cup \{\ell_4, \ell_2, \ell_5\}\ e_2 = \{\ell_1, \ell_3, \ell_4\} \cup \{\ell_2, \ell_5\}.$$

In order to compare the two possibilities then, we evaluate the phylogenetic invariants on the boundary distribution obtained from the data of SSWL variables for the five languages, distributed in the leaves of the tree in one of the two ways described above, and we compute estimates of the Euclidean distance function.

6.2 Syntactic Structures and Boundary Distributions

One of the main problems with the SSWL database is that the binary variables of syntactic structures are very non-uniformly mapped across languages. In order to use the data for phylogenetic reconstruction, it is necessary

P	[Tocharian A, Hittite, Albanian, Armenian, A.Greek]
01	[1,1,1,1,1]
06	[1,1,0,1,1]
11	[1,0,1,1,1]
12	[1,1,1,1,1]
13	[1,1,0,1,1]
15	[1,1,1,1,1]
17	[1,1,1,1,1]
19	[1,1,0,1,1]
21	[1,1,0,1,1]
A01	[1,1,1,0,1]
A02	[1,1,1,0,1]
Neg 01	[1,1,1,1,1]
Neg 03	[0,0,0,1,0]
Neg 04	[0,0,0,0,0]
Neg 07	[0,0,0,0,0]
Neg 08	[0,0,0,0,0]
Neg 09	[0,0,0,0,0]
Neg 10	[0,0,0,0,0]
Neg 12	[0,0,0,0,0]
Neg 13	[0,0,0,0,0]
Neg 14	[0,0,0,0,0]
Order N3 01	[1,1,1,1,1]

Fig. 2 The SSWL syntactic parameters P that are completely mapped for the set languages Tocharian A, Hittite, Albanian, Armenian, Ancient Greek, and their values on each language

to restrict to only those variables that are completely mapped for all the languages considered. In our present case, some of the languages are very poorly mapped in the SSWL database: Tocharian A is only 19% mapped, Tocharian B 18%, Hittite is 32% mapped, Albanian 69%, Armenian 89% and (Ancient) Greek is also 89% mapped. Moreover, not all the 29 binary syntactic variables that are mapped for Tocharian A are also among the variables mapped for Hittite. This reduces the list of syntactic variables that are completely mapped for all five of these languages to a total of only 22 variables. The variables (listed with the name used in the SSWL database) and the resulting values are given in the table in Fig. 2. Based on these data, the boundary distribution for the two cases considered above is given by the following. In the first case the frequencies are given by

$$p_{00000} = 4/11,\ p_{11111} = 3/11,\ p_{11101} = 2/11,$$
$$p_{11011} = 1/22,\ p_{10111} = 1/11,\ p_{01000} = 1/22$$

with $p_{i_1,\dots,i_5} = 0$ for all the remaining binary vectors in $\{0, 1\}^5$. In the second case we have frequencies

$$p_{00000} = 4/11,\ p_{11111} = 3/11,\ p_{11011} = 2/11,$$
$$p_{10111} = 1/22,\ p_{11101} = 1/11,\ p_{00010} = 1/22$$

with $p_{i_1,\dots,i_5} = 0$ for all the remaining binary vectors in $\{0, 1\}^5$.

For the first case, the flattening matrices evaluated at the boundary distribution P give the matrices

$$\mathrm{Flat}_{e_1,T_1} = \begin{pmatrix} \frac{4}{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{11} \\ 0 & 0 & 0 & \frac{1}{22} & 0 & \frac{2}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

$$\mathrm{Flat}_{e_2,T_1} = \begin{pmatrix} \frac{4}{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{11} \\ 0 & 0 & 0 & \frac{1}{22} \\ 0 & \frac{2}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

For the second case, on the other hand, we obtain the matrices

$$\mathrm{Flat}_{e_1,T_2} = \begin{pmatrix} \frac{4}{11} & 0 & \frac{1}{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{22} \\ 0 & 0 & 0 & \frac{2}{11} & 0 & \frac{1}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

$$\mathrm{Flat}_{e_2,T_2} = \begin{pmatrix} \frac{4}{11} & 0 & \frac{1}{22} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{22} \\ 0 & 0 & 0 & \frac{2}{11} \\ 0 & \frac{1}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

6.3 Phylogenetic Invariants

The evaluation of the phylogenetic invariants on these two boundary distributions by evaluating the 3×3 minors of the matrices above gives

(1) For the Gray-Atkins tree T_1:

$$\|\Phi_{T_1}(P)\|_{\ell^\infty} = \max_{\phi \in \mathcal{D}^{(3)}_{e_1,T_1} \cup \mathcal{D}^{(3)}_{e_2,T_1}} |\phi(P)| = \frac{8}{1331}$$

$$\|\Phi_{T_1}(P)\|_{\ell^1} = \sum_{\phi \in \mathcal{D}^{(3)}_{e_1,T_1} \cup \mathcal{D}^{(3)}_{e_2,T_1}} |\phi(P)| = \frac{61}{2662}$$

(2) For the Ringe–Warnow–Taylor tree T_2:

$$\|\Phi_{T_1}(P)\|_{\ell^\infty} = \max_{\phi \in \mathcal{D}^{(3)}_{e_1,T_2} \cup \mathcal{D}^{(3)}_{e_2,T_2}} |\phi(P)| = \frac{8}{1331}$$

$$\|\Phi_{T_1}(P)\|_{\ell^1} = \sum_{\phi \in \mathcal{D}^{(3)}_{e_1,T_2} \cup \mathcal{D}^{(3)}_{e_2,T_2}} |\phi(P)| = \frac{18}{1331}$$

On the basis of this naive test of evaluation of the phylogenetic invariants, the ℓ^∞ norm does not distinguish the two trees while the ℓ^1 norm prefers the Ringe–Warnow–Taylor tree T_2. We show below that this preference is also confirmed by an estimation of the Euclidean distance.

6.4 Estimate of the Euclidean Distance Function

In this case, in order to obtain a lower bound estimate of the Euclidean distance for the two trees T_1 and T_2, we compute the distances

$$D_{1,1} = \mathrm{dist}(\mathrm{Flat}_{e_1,T_1}(P), \mathcal{D}_2(4,8)), \quad D_{1,2} = \mathrm{dist}(\mathrm{Flat}_{e_2,T_2}(P), \mathcal{D}_2(8,4))$$

with the Euclidean distance estimate of T_1 given by $L_1 = \max\{D_{1,1}, D_{1,2}\}$ and

$$D_{2,1} = \mathrm{dist}(\mathrm{Flat}_{e_1,T_2}(P), \mathcal{D}_2(4,8)), \quad D_{2,1} = \mathrm{dist}(\mathrm{Flat}_{e_2,T_2}(P), \mathcal{D}_2(8,4))$$

with the Euclidean distance estimate of T_2 given by $L_2 = \max\{D_{2,1}, D_{2,2}\}$.

The computation of the singular values $\Sigma = (\sigma_1, \ldots, \sigma_4)$ of the flattening matrices $\mathrm{Flat}_{e_i,T_j}(P)$ gives

$$\Sigma(\mathrm{Flat}_{e_1,T_1}(P)) = \mathrm{diag}(0.3664662612, 0.3394847389, 0.5018672314 \times 10^{-1}, 0)$$
$$\Sigma(\mathrm{Flat}_{e_2,T_1}(P)) = \mathrm{diag}(0.3664662612, 0.3388120907, 0.5454321492 \times 10^{-1}, 0)$$
$$\Sigma(\mathrm{Flat}_{e_1,T_2}(P)) = \mathrm{diag}(0.3664662613, 0.3421098124, 0.2700872640 \times 10^{-1}, 0)$$
$$\Sigma(\mathrm{Flat}_{e_2,T_2}(P)) = \mathrm{diag}(0.3664662613, 0.3394847388, 0.5018672301 \times 10^{-1}, 0).$$

Since the last singular value is always zero, the Euclidean distances are given by the σ_3 value

$$D_{1,1} = 0.5018672314 \times 10^{-1}, \quad D_{1,2} = 0.5454321492 \times 10^{-1},$$
$$D_{2,1} = 0.2700872640 \times 10^{-1}, \quad D_{2,1} = 0.5018672301 \times 10^{-1}$$

This gives $L_1 = 0.5454321492 \times 10^{-1}$ and $L_2 = 0.5018672301 \times 10^{-1}$.

Thus, the Euclidean distance estimate also favors the Ringe–Warnow–Taylor tree T_2 over the Gray-Atkins tree T_1. The fact that there are very few parameters that are mapped (at present time) for all of these languages in the SSWL database, and that these parameters largely agree on this set of languages, however make this analysis not fully reliable. A more extensive set of syntactic data for these languages would be needed to confirm whether the phylogenetic reconstruction based on syntactic data and the algebro-geometric method is reliable.

7 Towards Larger Phylogenetic Trees: Grafting

As we have seen in the previous sections, Phylogenetic Algebraic Geometry is a procedure that associates to a given language family $\mathcal{L} = \{\ell_1, \ldots, \ell_n\}$ an algebraic variety $Y = Y(\mathcal{L}, P)$ constructed on the basis of the syntactic variables (listed in the distribution P).

A possible geometric viewpoint on comparative historical linguistics can then be developed, by considering the geometry of the varieties $Y(\mathcal{L}, P)$ for different language families. This contains more information than the topology of the tree by itself, in the sense that one can, for example, look more specifically for the position of the point P on the variety. The point P contains precise information on how the binary syntactic variables change across the languages in the family. For example, in the case of the six Germanic languages in the set $\mathcal{S}_1(G)$, we see from our table of occurrences that only very few possibilities for the binary vector $(i_1, \ldots, i_6)$ occur for these six languages. We also see that, apart from the cases where the value of a syntactic variable agrees in all six languages (40 occurrences where the feature is not expressed, and 22 where it is), we find that it is more likely for Icelandic to have a feature that differs from the other languages in the group (4 occurrences of $(0, 0, 0, 0, 1, 0)$ of lacking a features the others have and 3 occurrences of $(1, 1, 1, 1, 0, 1)$ for having a feature that the others lack). Thus,

the location of the point P on the variety contains information that is related to the spreading of syntactic features across the language family considered. This geometric way of thinking may be compared with the coding theory approach of [30,46] to measuring the spread of syntactic features across a language family.

As we have seen in the example discussed above of a small set of Germanic languages, as well as in the examples with Romance and Slavic languages, the use of SSWL data is suitable for phylogenetic reconstruction, provided only the subset of the completely mapped syntactic variables (for the given set of languages) is used and the candidate phylogenetic trees are selected through the computation of phylogenetic invariants, and their evaluation at the boundary distribution determined by the syntactic variables.

This method works very well for small trees and for a set of languages that is well mapped in the available databases (with enough binary syntactic variables that are mapped for all the languages in the given set). However, one then needs a way to combine phylogenetic trees of smaller subfamilies into those of larger families.

We give a very brief sketch of how this procedure can be articulated in terms of Phylogenetic Algebraic Geometry, and we refer the readers to §5–8 of [1] for more details. Although we do not need to use this method directly in the present paper, we mention this for completeness, since it is a natural question how to proceed towards larger trees. Given two binary trees T' and T'', respectively with n and m leaves, the grafting $T = T' \star_\ell T''$ at a leaf ℓ is the binary tree obtained by gluing together a leaf of T' with marking ℓ to a leaf of T'' with the same marking. The resulting tree T has $n + m - 2$ leaves. It is shown in [1] how the phylogenetic invariants of T depend on the invariants of T' and T''. Consider the maps $\Phi_{T'}$ and $\Phi_{T''}$, defined as in (2.2) using (2.1), with values in $\mathbb{C}^{2^n}$ and $\mathbb{C}^{2^m}$, respectively. We identify $\mathbb{C}^{2^n} = \mathbb{C}^{2^{n-1}} \otimes \mathbb{C}^2$, where the last binary variable corresponds to the leaf ℓ. We then identify the affine space $\mathbb{C}^{2^{n-1}} \otimes \mathbb{C}^2 \simeq \mathrm{Hom}(\mathbb{C}^{2^{n-1}\vee}, \mathbb{C}^2)$ with the space of matrices $M_{2^{n-1}\times 2}(\mathbb{C})$, and similarly with $\mathbb{C}^{2^m} \simeq M_{2\times 2^{m-1}}(\mathbb{C})$. One then defines $\Phi_T = \Phi_{T'} \star \Phi_{T''}$ as the matrix product of the elements in the range of $\Phi_{T'}$, seen as matrices in $M_{2^{n-1}\times 2}(\mathbb{C})$ with the elements in the range of $\Phi_{T''}$, seen as matrices in $M_{2\times 2^{m-1}}(\mathbb{C})$. This results in a matrix in $M_{2^{n-1}\times 2^{m-1}}(\mathbb{C})$), which gives a map Ψ_T with values in $\mathbb{C}^{n+m-2}$. The domain variables of Ψ_T are obtained as follows. For those edges of T not involved in the grafting operation, we define the 2×2 matrices M^e to be the same as those originally associated to the edges of T' or T'', respectively. For the edge of T' and the edge of T'' that are glued together in the grafting, we replace the respective matrices $M^{e'}$ and $M^{e''}$ by their product $M^e = M^{e'} M^{e''}$. Dually, as in (2.3), this determines the map Ψ_T of polynomial rings, whose kernel is the phylogenetic ideal of T. The closure in $\mathbb{C}^{n+m-2}$ of the image of Ψ_T is the phylogenetic algebraic variety of the grafted tree $T = T' \star_\ell T''$.

Suppose we are interested in the phylogenetic tree of a language family $\mathcal{L}$, for which we assume that we already know (from other linguistic input) a subdivision into several subfamilies $\mathcal{L} = \mathcal{L}_1 \cup \cdots \cup \mathcal{L}_N$. Suppose also that for the language families taken into consideration there are sufficient data available about the ancient languages. (This requirement will limit the applicability of the algorithm discussed here to families like the Indo-European, where significant amount of data about ancient languages is available.) We can then follow the following procedure to graft phylogenetic trees of the subfamilies $\mathcal{L}_k$ into a larger phylogenetic tree for the family $\mathcal{L}$. For the procedure described here we need to assume that one knows a priori (via historical linguistic information) that the members of the subfamilies $\mathcal{L}_k$ should remain together in a clade of the grafted tree.

(1) For each subfamily $\mathcal{L}_k = \{\ell_{k,1}, \ldots, \ell_{k,n_k}\}$, consider the list of SSWL data that are completely mapped for all the languages $\ell_{k,j}$ in the subfamily $\mathcal{L}_k$.
(2) On the basis of that set of binary syntactic variables, a preferred candidate phylogenetic tree T_k is constructed based on the method illustrated above in the example of the Germanic languages.
(3) Use the procedure discussed in Sect. 3.5 above to identify the best location of the root vertex for each tree T_k, and regard each tree T_k as a tree with $n_k + 1$ leaves, including one leaf attached to the root vertex.
(4) Let $\{\lambda_1, \ldots, \lambda_N\}$ be the ancient languages located at the root vertex of each tree $T_1, \ldots, T_N$. Consider the list of SSWL parameters that are completely mapped for all the ancient languages λ_k.
(5) On the basis of that set of binary syntactic variables, select preferred candidate phylogenetic tree T with N leaves, by evaluating the phylogenetic invariants of these trees on the boundary distribution given by this set of binary syntactic variables.

(6) Graft the best candidate tree T to the trees T_k by gluing the leaf λ_k of T to the root of T_k.
(7) The phylogenetic invariants of the resulting grafted tree $T' = T \star_{k=1}^{N} T_k$ can be computed with the grafting procedure of [1] described above and evaluation at the boundary distribution given by the leaves $\{\ell_{k,j} \mid j = 1, \dots, n_k,\ k = 1, \dots, N\}$ of T' (coming from the smaller set of syntactic variables that are completely mapped for all the $\ell_{k,j}$) can confirm the selected tree topology T'.

The advantage of this procedure is that it is going to work even in the absence of a sufficient number of binary syntactic variables in the SSWL database that are completely mapped for all of the languages $\ell_{k,j}$ at the same time, provided there are enough for each subset $\mathcal{L}_k$ and for the λ_k. In cases where the number of variables that are completely mapped for all the $\ell_{k,j}$ is significantly smaller compared to those that are mapped within each group, the last test on the tree T' becomes less significant. This method also has the advantage that one works with the smaller subtrees T_k and T, rather than with the bigger tree given by their grafting, so that the computations of phylogenetic invariants is more tractable.

In the case of language families where one does not have syntactic data of ancient languages available, one can still adapt the procedure described above, provided there is a reasonable number of SSWL variables that are completely mapped for all the languages $\ell_{k,j}$ in $\mathcal{L}$. One can proceed as follows.

(1) For each subfamily $\mathcal{L}_k = \{\ell_{k,1}, \dots, \ell_{k,n_k}\}$, consider the list of data that are completely mapped for all the languages $\ell_{k,j}$ in the subfamily $\mathcal{L}_k$.
(2) On the basis of that set of binary syntactic variables, a preferred candidate phylogenetic tree T_k is constructed based on the method illustrated above in the example of the Germanic languages.
(3) Consider all possible choices of a root vertex for each T_k (there are as many choices as the number of internal edges of T_k).
(4) Consider all the possible candidate tree topologies T with N leaves.
(5) For each choice of a root vertex in each T_k graft a choice of T to the give roots of the trees T_k to obtain a candidate tree $T' = T \star_{k=1}^{N} T_k$.
(6) Compute the phylogenetic invariants of $T' = T \star_{k=1}^{N} T_k$ using the procedure of [1] recalled above.
(7) Evaluate the phylogenetic invariants of each candidate T' on the boundary distribution determined by the binary syntactic variables that are completely mapped for all the languages $\{\ell_{k,j} \mid j = 1, \dots, n_k,\ k = 1, \dots, N\}$, to select the best candidate among the T'.

There are serious computational limitations to this procedure, however, because of how fast the number of trees on N leaves grows. While the grafting procedure discussed above makes it possible to work with smaller trees and then consider the problem of grafting them into a larger tree, this would still only work computationally for small size trees, and cannot be expected to handle, for example, the entire set of languages recorded in the SSWL database.

8 Modifying the Setting to Account for Syntactic Relations

In a followup to this paper, based on the ongoing analysis of [34], we will discuss how to adjust these phylogenetic models to incorporate deviations from the assumption that the syntactic parameters are i.i.d. random variables evolving according to the same Markov model on a tree.

Indeed, we know from various data analysis of the syntactic variables, including topological data analysis [41,42], methods of coding theory [46], and recoverability in Kanerva networks [38], that the syntactic parameters are certainly not i.i.d. variables. Thus, it is likely that some discrepancies we observed in this paper, in the application of the Phylogenetic Algebraic Geometry method (for example in the case of the Romance languages or the early Indo-European languages where the tree selected by the Euclidean distance is not the same as the tree favored by the phylogenetic invariants) may be an effect of the use of this overly simplified assumption.

The approach we plan to follow to at least partially correct for this problem, is to modify the boundary distribution on the tree by attaching to the different syntactic parameters a weight that comes from some measure of its dependence

from other parameters, in such a way that parameters that are more likely to be dependent variables according to one of these tests will weight less in the boundary distribution than parameters that are more likely to be truly statistically independent variables.

The main idea on how to achieve this goal is to modify the boundary distribution P by counting occurrences $n_{i_1,\dots,i_n}$ of parameter values $(i_1,\dots,i_n)$ at the n leaves of the tree by introducing weights for different parameters that measure their degree of independence. An example of such a weight would be the degree of recoverability in a Kanerva network, as in [38], or a computation of clustering coefficients as in [34].

This means that, instead of assigning to a given binary vector $(i_1,\dots,i_n)$ the frequency

$$p_{i_1,\dots,i_n} = \frac{n_{i_1,\dots,i_n}}{N}$$

with N total number of parameters and $n_{i_1,\dots,i_n}$ number of parameters that have values $(i_1,\dots,i_n)$ on the n languages at the leaves of the tree, we replace this by a new distribution

$$p'_{i_1,\dots,i_n} = Z^{-1} \sum_{r=1}^{n_{i_1,\dots,i_n}} w(\pi_r)$$

where for a syntactic parameter π the weight $w(\pi)$ measures the degree of independence of π, for example with $w(\pi)$ close to 1 the more π can be regarded as an independent variable and close to 0 the more π is recoverable from the other variables, with Z a normalization factor so that $p'_{i_1,\dots,i_n}$ is again a probability distribution.

With this new boundary distribution P' we will recompute the Euclidean distances of the flattening matrices $\mathrm{Flat}_{e,T}(P')$ from the varieties $\mathcal{D}_2(a,b)$ by computing the singular values $(\sigma_1,\dots,\sigma_a)$ of $\mathrm{Flat}_{e,T}(P')$ and computing the square-distance as $\sigma_3^2+\cdots+\sigma_a^2$, and compare the new distances obtained in this way with those of the original boundary distribution P.

Results on this approach will be presented in forthcoming work.

Acknowledgements The first and second author were partially supported by a Summer Undergraduate Research Fellowship at Caltech. The last author is partially supported by NSF Grant DMS-1707882, NSERC Discovery Grant RGPIN-2018-04937, Accelerator Supplement Grant RGPAS-2018-522593, and by the Perimeter Institute for Theoretical Physics. We are very grateful to the two anonymous referees for many very useful comments, corrections, and suggestions that greatly improved the paper.

Appendix A: SSWL Syntactic Variables of the Set $\mathcal{S}_1(G)$ of Germanic Languages

We list here the 90 binary syntactic variables of the SSWL database that are completely mapped for the six Germanic languages ℓ_1 = Dutch, ℓ_2 = German, ℓ_3 = English, ℓ_4 = Faroese, ℓ_5 = Icelandic, ℓ_6 = Swedish. The column on the left in the tables lists the SSWL parameters P as labeled in the database.

P	$[\ell_1,\ell_2,\ell_3,\ell_4,\ell_5,\ell_6]$
01	[1,1,1,1,1,1]
02	[0,0,0,0,0,0]
03	[1,1,1,1,1,1]
04	[1,1,0,0,0,0]
05	[1,1,1,1,1,1]
06	[1,1,0,0,0,0]
07	[0,0,0,0,0,0]
08	[0,0,0,0,0,0]
09	[0,0,0,0,0,0]
10	[0,0,0,0,0,0]
11	[1,1,1,1,1,1]
12	[1,0,0,0,0,0]
13	[1,1,1,1,1,1]
14	[0,0,1,1,1,0]
15	[1,1,1,1,1,1]
16	[0,0,0,0,1,0]
17	[1,1,1,1,1,1]
18	[0,0,0,0,1,0]
19	[1,1,1,1,0,1]
20	[1,1,1,1,1,1]
21	[1,1,1,1,0,1]
22	[0,0,0,1,1,0]
A01	[1,1,0,1,1,1]
A02	[0,0,0,1,1,1]
A03	[1,1,1,1,1,1]
A04	[1,1,1,0,0,1]
Aux Sel 01	[1,1,0,1,1,0]
C01	[1,1,1,1,1,1]
C02	[0,0,0,0,0,0]
C03	[1,1,1,1,1,1]
C04	[0,0,0,0,0,0]
EE	[1,1,1,1,1,0]

P	$[\ell_1,\ell_2,\ell_3,\ell_4,\ell_5,\ell_6]$
N2 01	[1,1,1,1,1,1]
N2 02	[0,0,0,1,0,0]
N2 03	[1,1,1,1,0,1]
N2 04	[0,0,0,0,1,0]
N2 05	[1,1,1,1,1,1]
N2 06	[1,1,1,1,1,1]
N2 07	[0,0,0,0,0,0]
N2 08	[0,0,0,0,0,0]
N2 09	[0,0,0,0,0,0]
N2 10	[0,0,0,0,1,0]
N2 11	[0,0,0,0,1,1]
Neg 01	[1,1,1,0,0,1]
Neg 02	[1,1,1,1,1,1]
Neg 03	[0,0,0,0,0,0]
Neg 04	[0,0,1,0,0,0]
Neg 05	[0,0,0,0,0,0]
Neg 06	[0,0,0,0,0,0]
Neg 07	[0,0,0,0,0,0]
Neg 08	[0,0,0,0,0,0]
Neg 09	[0,0,0,0,0,0]
Neg 10	[0,0,0,0,0,0]
Neg 11	[0,0,0,0,0,0]
Neg 12	[0,0,0,0,0,0]
Neg 13	[0,1,0,0,0,0]
Neg 14	[0,0,0,0,0,0]

P	$[\ell_1,\ell_2,\ell_3,\ell_4,\ell_5,\ell_6]$
Order N3 01	[1,1,1,1,1,1]
Order N3 02	[0,0,1,0,0,0]
Order N3 03	[0,0,0,0,0,0]
Order N3 04	[0,0,0,0,0,0]
Order N3 05	[0,0,0,0,0,0]
Order N3 06	[0,0,0,0,0,0]
Order N3 07	[1,1,1,1,1,1]
Order N3 09	[0,0,0,0,0,0]
Order N3 10	[0,0,0,0,0,0]
Order N3 11	[0,0,0,0,0,0]
Order N3 12	[0,0,0,0,0,0]
Q01	[0,0,0,0,0,0]
Q02	[0,0,0,0,0,0]
Q03	[0,0,0,0,0,0]
Q05	[1,1,0,1,1,1]
Q06	[1,1,1,1,1,1]
Q07	[0,0,0,0,0,0]
Q08	[1,1,1,1,1,1]
Q09	[0,0,0,0,0,0]
Q10	[0,0,0,0,0,0]
Q11	[0,0,0,0,0,0]
Q12	[0,0,0,0,0,0]
Q13	[0,0,0,0,0,0]
Q14	[0,0,0,0,0,0]
Q15	[0,0,0,0,0,0]
Q16NEGA	[1,1,1,1,1,1]
Q17NEGA	[0,0,0,1,0,0]
Q18NEGA	[0,0,0,0,0,0]
Q20ANegQ	[1,1,1,1,1,1]
Q21ANegQ	[0,1,0,1,1,1]
Q22ANegQ	[1,0,0,0,0,0]
V2 01	[1,1,0,1,1,1]
V2 02	[1,1,1,1,1,1]

Appendix B: SSWL Syntactic Variables of the Set $\mathcal{S}_2(G)$ of Germanic Languages

We list here the 90 binary syntactic variables of the SSWL database that are completely mapped for the seven Germanic languages ℓ_1 = Norwegian, ℓ_2 = Danish, ℓ_3 = Gothic, ℓ_4 = Old English, ℓ_5 = Icelandic, ℓ_6 = English, ℓ_7 = German. The column on the left in the tables lists the SSWL parameters P as labeled in the database.

P	$[\ell_1,\ell_2,\ell_3,\ell_4,\ell_5,\ell_6,\ell_7]$
01	[1,1,1,1,1,1,1],
03	[1,1,1,1,1,1,1],
04	[0,0,1,1,0,0,1],
05	[1,1,1,1,1,1,1],
06	[0,0,1,1,0,0,1],
07	[0,0,1,0,0,0,0],
08	[0,0,0,0,0,0,0],
09	[0,0,0,0,0,0,0],
10	[0,0,0,0,0,0,0],
11	[1,1,1,1,1,1,1],
12	[1,1,0,1,0,0,0],
13	[1,1,1,1,1,1,1],
14	[0,0,1,1,1,1,0],
15	[1,1,1,1,1,1,1],
16	[0,0,1,1,1,0,0],
17	[1,1,1,1,1,1,1],
18	[0,0,1,0,1,0,0],
19	[1,1,0,1,0,1,1],
20	[1,1,1,1,1,1,1],
21	[1,1,0,1,0,1,1],
22	[1,0,1,1,1,0,0],
A01	[1,1,1,1,1,0,1],
A02	[1,1,1,1,1,0,0],
A03	[1,1,1,1,1,1,1],
A04	[1,1,1,1,0,1,1],
Aux Sel 01	[1,1,0,1,1,0,1],
C01	[1,1,1,1,1,1,1],
C02	[0,0,0,0,0,0,0],
C03	[1,1,1,1,1,1,1],
C04	[0,0,0,0,0,0,0],

P	$[\ell_1,\ell_2,\ell_3,\ell_4,\ell_5,\ell_6,\ell_7]$
N2 01	[1,1,1,1,1,1,1],
N2 02	[0,0,1,0,0,0,0],
N2 03	[1,1,1,1,0,1,1],
N2 05	[1,1,1,1,1,1,1],
N2 06	[1,1,1,1,1,1,1],
N2 08	[0,0,0,0,0,0,0],
N2 09	[0,0,0,0,0,0,0],
N2 10	[0,0,0,0,1,0,0],
N2 11	[1,1,0,0,1,0,0],
Neg 01	[1,1,1,1,0,1,1],
Neg 02	[1,1,0,0,1,1,1],
Neg 03	[0,0,0,0,0,0,0],
Neg 04	[0,0,0,0,0,1,0],
Neg 05	[0,0,0,0,0,0,0],
Neg 06	[0,0,0,0,0,0,0],
Neg 07	[0,0,0,0,0,0,0],
Neg 08	[0,0,0,0,0,0,0],
Neg 09	[0,0,0,0,0,0,0],
Neg 10	[0,0,0,0,0,0,0],
Neg 11	[0,0,0,1,0,0,0],
Neg 12	[0,0,0,0,0,0,0],
Neg 13	[0,0,0,0,0,0,1],
Neg 14	[0,0,0,0,0,0,0]

P	$[\ell_1,\ell_2,\ell_3,\ell_4,\ell_5,\ell_6,\ell_7]$
Order N3 04	[0,0,1,1,0,0,0],
Order N3 07	[1,1,1,1,1,1,1],
Order N3 08	[0,0,0,1,0,0,0],
Q01	[0,0,0,0,0,0,0],
Q02	[0,0,0,0,0,0,0],
Q03	[0,0,1,0,0,0,0],
Q06	[1,1,0,1,1,1,1],
Q07	[0,0,0,0,0,0,0],
Q08	[1,1,1,1,1,1,1],
Q10	[0,0,0,0,0,0,0],
Q11	[0,0,0,0,0,0,0],
Q12	[0,0,0,0,0,0,0],
Q13	[0,0,0,0,0,0,0]
Q15	[0,0,0,0,0,0,0],
Q17NEGA	[0,0,0,0,0,0,0],
Q18NEGA	[0,0,0,0,0,0,0],

Appendix C: Flattening Matrices F_5 and F_6

The flattening matrices of (3.1) (written in transpose form for convenience) for the T_5 and T_6 trees, in the case of the Longobardi data are given by the following:

$$F_5^t = \begin{pmatrix}
\frac{4}{7} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{42} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{42} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & \frac{1}{42} & 0 & \frac{1}{42} \\
0 & 0 & 0 & \frac{1}{42} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{42} & \frac{2}{7}
\end{pmatrix} \qquad F_6^t = \begin{pmatrix}
\frac{4}{7} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{42} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{42} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & \frac{1}{42} & 0 & 0 \\
0 & 0 & 0 & \frac{1}{42} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & \frac{1}{42} & \frac{1}{42} & \frac{2}{7}
\end{pmatrix}$$

The same flattening matrices of (3.1) for the SSWL data are given by the following.

$$F_5^t = \begin{pmatrix}
\frac{13}{34} & \frac{1}{68} & \frac{1}{34} & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{68} & 0 \\
\frac{3}{68} & \frac{1}{68} & \frac{1}{68} & \frac{1}{68} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{68} \\
0 & 0 & 0 & \frac{1}{68} \\
\frac{1}{68} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{68} \\
0 & 0 & \frac{1}{34} & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{68} \\
\frac{1}{68} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{68} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & \frac{1}{68} & \frac{1}{34} & \frac{1}{68} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & \frac{3}{68} & \frac{4}{17}
\end{pmatrix} \quad F_6^t = \begin{pmatrix}
\frac{13}{34} & \frac{1}{68} & \frac{3}{68} & \frac{1}{68} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\frac{1}{34} & 0 & \frac{1}{68} & \frac{1}{68} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{68} \\
\frac{1}{68} & 0 & 0 & \frac{1}{68} \\
\frac{1}{68} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{34} & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & \frac{1}{68} & 0 & \frac{1}{68} \\
\frac{1}{68} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{68} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & \frac{1}{68} & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\frac{1}{34} & \frac{1}{68} & \frac{3}{68} & \frac{4}{17}
\end{pmatrix}$$

Appendix D: List of LanGeLin Syntactic Parameters

FGP	Gramm. person	GSI	Grammaticalised inalienability
FGM	Gramm. Case	ALP	Alienable possession
FPC	Gramm. perception	GST	Grammaticalised Genitive
FGT	Gramm. temporality	GEI	Genitive inversion
FGN	Gramm. number	GNR	Non-referential head marking
GCO	Gramm. collective number	STC	Structured cardinals
PLS	Plurality spreading	GPC	Gender polarity cardinals
FND	Number in D	PMN	Personal marking on numerals
FSN	Feature spread on N	CQU	Cardinal quantifiers
FNN	Number in N	PCA	Number spread through cardinal adjectives
SGE	Semantic gender	PSC	Number spread from cardinal quantifiers
FGG	Gramm. gender	RHM	Head-markong on Rel
CGB	Unbounded sg N	FRC	Verbal relative clauses
DGR	Gramm. amount	NRC	Nominalized relative clause
DGP	Gramm. text anaphora	NOR	NP over verbal rel clauses/adpos gen
CGR	Strong amount	AER	Relative extrap.
NSD	Strong person	ARR	Free reduced rel
FVP	Variable person	DOR	def on relatives
DGD	Gramm. distality	NOD	NP over D
DPQ	Free null partitive Q	NOP	NP over non-genitive arguments
DCN	Article-checking N	PNP	P over complement
DNN	Null-N-licensing art	NPP	N-raising with obl. pied-piping
DIN	D-controlled infl. on N	NGO	N over GenO
FGC	Gramm. classifier	NOA	N over As
DBC	Strong classifier	NM2	N over M2 As
XCN	Conjugated nouns	NM1	N over M1 As
GSC	c-selection	EAF	Fronted high As
NOE	N over ext. arg.	NON	N over numerals
HMP	NP-heading modifier	FPO	Feature spread to genitive postpositions
AST	Structured APs	ACM	Class MOD
FFS	Feature spread to struct. APs	DOA	def on all +N
ADI	D-controlled infl. on A	NEX	Gramm. expletive article
DMP	def matching pron. poss.	NCL	Clitic poss.
DMG	def matching genitives	PDC	Article-checking poss.
GCN	Poss^o-checking N	ACL	Enclitic poss. on As
GFN	Gen-feature spread to Poss^o	APO	Adjectival poss.
GAL	Dependent Case in NP	WAP	Wackernagel adjectival poss.
GUN	Uniform Gen	AGE	Adjectival Gen
EZ1	Generalized linker	OPK	Obligatory possessive with kinship noun
EZ2	Non-clausal linker	TSP	Split deictic demonstratives
EZ3	Non-genitive linker	TSD	Split demonstratives
GAD	Adpositional Gen	TAD	Adjectival demonstratives
GFO	GenO	TDC	Article-checking demonstratives
PGO	Partial GenO	TLC	Loc-checking demonstratives
GFS	GenS	TNL	NP over Loc
GIT	Genitive-licensing iterator		

References

1. Allman, E., Rhodes, J.: Phylogenetic ideals and varieties for general Markov models. Adv. Appl. Math. **40**, 127–148 (2008)
2. Anthony, D.W., Ringe, D.: The Indo-European homeland from linguistic and archaeological perspectives. Annu. Rev. Linguist. **1**, 199–219 (2015)
3. Baker, M.: The Atoms of Language. Basic Books, New York (2001)
4. Barbançon, F., Evans, S.N., Nakhleh, L., Ringe, D., Warnow, T.: An experimental study comparing linguistic phylogenetic reconstruction methods. Diachronica **30**(2), 143–170 (2013)
5. Bocci, C.: Topics in phylogenetic algebraic geometry. Expo. Math. **25**, 235–259 (2007)
6. Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S.J., Alekseyenko, A.V., Drummond, A.J., Gray, R.D., Suchard, M.A., Atkinson, Q.D.: Mapping the origins and expansion of the Indo-European language family. Science **337**, 957–960 (2012)

7. Bruns, W., Vetter, U.: Determinantal Rings. Lecture Notes in Mathematics, vol. 1327. Springer, Berlin (1988)
8. Casanellas, M., Fernández-Sánchez, J.: Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees. Mol. Biol. Evol. **24**(1), 288–293 (2007)
9. Cartwright, D., Häbich, M., Sturmfels, B., Werner, A.: Mustafin varieties. Selecta Math. (N.S.) **17**(4), 757–793 (2011)
10. Chomsky, N.: Lectures on Government and Binding. Foris Publications, Dordrecht (1982)
11. Chomsky, N.: The Minimalist Program, 20th, Anniversary MIT Press (2015)
12. Chomsky, N., Lasnik, H.: The theory of Principles and Parameters. In: Syntax: An International Handbook of Contemporary Research, pp. 506–569, de Gruyter, (1993)
13. Draisma, J., Horobeţ, E., Ottaviani, G., Sturmfels, B., Thomas, R.: The Euclidean distance degree of an algebraic variety. Found. Comput. Math. **16**(1), 99–149 (2016)
14. Eriksson, N.: Using invariants for phylogenetic tree construction. In: Emerging Applications of Algebraic Geometry, IMA Volumes in Mathematics and Its Applications, vol. 149, pp. 89–108. Springer (2009)
15. Eriksson, N., Ranestad, K., Sturmfels, B., Sullivant, S.: Phylogenetic Algebraic Geometry. In: Projective Varieties with Unexpected Properties, pp. 237–255. Walter de Gruyter (2005)
16. Forster, P., Renfrew, C.: Phylogenetic Methods and the Prehistory of Language. McDonald Institute Monographs, Cambridge (2006)
17. Gakkhar, S., Marcolli, M.: Syntactic structures and the general Markov model, in preparation
18. Gray, R.D., Atkinson, Q.D.: Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature **426**(6965), 435–439 (2003)
19. Gusfield, D.: Recombinatorics. MIT Press, Cambridge (2014)
20. Harris, J.: Algebraic Geometry. Springer, Berlin (2013)
21. Kazakov, D., Cordoni, G., Algahtani, E., Ceolin, A., Irimia, M., Kim, S.S., Michelioudakis, D., Radkevich, N., Guardiano, C., Longobardi, G.: Learning implicational models of Universal Grammar parameters. In: EVOLANG XII, pp. 16–19 April 2018, Torun, Poland
22. Karimi, S., Piattelli-Palmarini M. (eds.): Special Issue on Parameters, Linguistic Analysis, vol. 41, No. 3–4 (2017)
23. Hauenstein, J., Rodriguez, J.I., Sturmfels, B.: Maximum likelihood for matrices with rank constraints. J. Algebr. Stat. **5**(1), 18–38 (2014)
24. Longobardi, G.: Principles, parameters, and schemata. A constructivist UG. Linguist. Anal. **41**(3–4), 517–556 (2017)
25. Longobardi, G.: A minimalist program for parametric linguistics? In: Broekhuis, H., Corver, N., Huybregts, M., Kleinhenz, U., Koster, J. (eds.) Organizing Grammar: Linguistic Studies for Henk van Riemsdijk, pp. 407–414. Mouton de Gruyter, Berlin (2005)
26. Longobardi, G.: Methods in parametric linguistics and cognitive history. Linguist. Var. Yearb. **3**, 101–138 (2003)
27. Longobardi, G., Guardiano, C.: Evidence for syntax as a signal of historical relatedness. Lingua **119**, 1679–1706 (2009)
28. Longobardi, G., Guardiano, C., Silvestri, G., Boattini, A., Ceolin, A.: Towards a syntactic phylogeny of modern Indo-European languages. J. Hist. Linguist. **3**(1), 122–152 (2013)
29. Longobardi, G., Buch, A., Ceolin, A., Ecay, A., Guardiano, C., Irimia, M., Michelioudakis, D., Radkevich, N., Jaeger, G.: Correlated evolution or not? phylogenetic linguistics with syntactic, cognacy, and phonetic data. In: Roberts, S.G. et al. (eds.) The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11), 2016 Online at http://evolang.org/neworleans/papers/162.html. (2016)
30. Marcolli, M.: Syntactic parameters and a coding theory perspective on entropy and complexity of language families. Entropy **18**(4), 110 (2016)
31. Mirsky, L.: Symmetric gauge functions and unitarily invariant norms. Q. J. Math. **11**, 1156–1159 (1966)
32. Murawaki, Y.: Continuous space representations of linguistic typology and their application to phylogenetic inference. In: Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pp. 324–334 (2015)
33. Nakhleh, L., Ringe, D., Warnow, T.: Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. Language **81**(2), 382–420 (2005)
34. Ortegaray, A., Berwick, R.C., Marcolli, M.: Heat kernel analysis of syntactic structures. arXiv:1803.09832, to appear in Mathematics in Computer Science
35. Pachter, L., Sturmfels, B.: The mathematics of phylogenomics. SIAM Rev. **49**(1), 3–31 (2007)
36. Pachter, L., Sturmfels, B.: Tropical geometry of statistical models. Proc. Natl. Acad. Sci. (PNAS) **101**(46), 16132–16137 (2004)
37. Pachter, L., Sturmfels, B.: Algebraic Statistics for Computational Biology. Cambridge University Press, Cambridge (2005)
38. 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, Lecture Notes in Computer Science, vol. 10589, pp. 265–272. Springer (2017)
39. Perelysvaig, A., Lewis, M.W.: The Indo-European Controversy: Facts and Fallacies in Historical Linguistics. Cambridge University Press, Cambridge (2015)
40. PHYLIP: http://evolution.genetics.washington.edu/phylip.html
41. Port, A., Gheorghita, I., Guth, D., Clark, J.M., Liang, C., Dasu, S., Marcolli, M.: Persistent topology of syntax. Math. Comput. Sci. **12**(1), 33–50 (2018)
42. Port, A., Karidi, T., Marcolli, M.: Topological analysis of syntactic structures. arXiv:1903.05181
43. Ringe, D., Warnow, T., Taylor, A.: Indo-European and computational cladistics. Trans. Philol. Soc. **100**, 59–129 (2002)
44. Rizzi, L.: On the format and locus of parameters: the role of morphosyntactic features. Linguist. Anal. **41**, 159–191 (2017)

45. Rusinko, J.P., Hipp, B.: Invariant based quartet puzzling. Algorithms Mol. Biol. **7**, 35 (2012)
46. Shu, K., Marcolli, M.: Syntactic structures and code parameters. Math. Comput. Sci. **11**(1), 79–90 (2017)
47. 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, pp. 417–441. Springer, Berlin (2018)
48. Siva, K., Tao, J., Marcolli, M.: Spin glass models of syntax and language evolution. Linguist. Anal. **41**(3–4), 559–608 (2017)
49. SSWL Database of Syntactic Parameters: http://sswl.railsplayground.net/
50. Sturmfels, B., Sullivant, S.: Toric ideals of phylogenetic invariants. J. Comput. Biol. **12**(2), 204–228 (2005)
51. Warnow, T.: Computational Phylogenetics. Cambridge University Press, Cambridge (2017)
52. Warnow, T., Evans, S.N., Ringe, D., Nakhleh, L.: Stochastic models of language evolution and an application to the Indo-European family of languages. Available at http://www.stat.berkeley.edu/users/evans/659.pdf