BRILL

LANGUAGE DYNAMICS AND CHANGE 8 (2018) 55–77

brill.com/ldc

Copystree

Gaming artificial phylogenies

Simone Pompei
Institute for Theoretical Physics, University of Cologne, Köln
simo.pompei@gmail.com

Vittorio Loreto
SONY Computer Science Laboratories, Paris; Sapienza University of Rome, Physics Deptartment, Rome; Complexity Science Hub Vienna, Vienna
vittorio.loreto@roma1.infn.it

Francesca Tria
Institute for Theoretical Physics, University of Cologne, Köln; Sapienza University of Rome, Physics Deptartment, Rome
fratrig@gmail.com

Abstract

The reconstruction of phylogenies of cultural artefacts represents an open problem that mixes theoretical and computational challenges. Existing benchmarks rely on simulated phylogenies, where hypotheses on the underlying evolutionary mechanisms are unavoidable, or on real data phylogenies, for which no true evolutionary history is known. Here we introduce a web-based game, *Copystree*, where users create phylogenies of manuscripts through successive copying actions in a fully monitored setup. While players enjoy the experience, Copystree allows to build artificial phylogenies whose evolutionary processes do not obey any predefined theoretical mechanisms, being generated instead with the unpredictability of human creativity. We present the analysis of the data gathered during the first set of experiments and use the artificial phylogenies gathered for a first test of existing phylogenetic algorithms.

Keywords

scientific gaming – phylogenetic reconstruction – stemmatics – language evolution

 | DOI: 10.1163/22105832-00801003

1 Introduction

The relationship between language change and biological evolution (Maynard Smith and Szathmary, 1997) has been investigated since the emergence of linguistics as a science in the nineteenth century, paralleling the emergence of evolutionary theory. The observation of languages changing had a documented influence on Darwin's thoughts. In *The Origin of Species* (Darwin, 1859), Darwin argued that our ability to order languages genealogically, despite their having changed and divided at different rates, allows us to think that the same can be done for species. And in *The Descent of Man* (Darwin, 1871), he noted, 'The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel.' These similarities have been further investigated and explored over time, and nowadays modern computational and mathematical tools of evolutionary analysis (Felsenstein, 2004; Gascuel, 2005), initially developed in biology, have been successfully applied in historical linguistics (Renfrew et al., 2000; Joseph and Janda, 2004; Wichmann and Grant, 2012; Tria et al., 2010a; Gray and Atkinson, 2003; Bryant et al., 2005; Pagel et al., 2007; Atkinson et al., 2008; Dunn et al., 2008; Gray et al., 2009).

Phylogenetic reconstruction, in particular, is the research field where this cross-fertilization has been especially fruitful. Although the application of quantitative studies in historical linguistics is not a novel topic, dating back to the 50's and the introduction of the lexicostatistics approach by Swadesh (1952, 1955), in the last decades we have witnessed an unprecedented number of computational and phylogenetic applications in this field. For instance, inferred language trees have been successfully used to evaluate evolutionary scenarios of human history (Gray and Atkinson, 2003; Bryant et al., 2005; Pagel et al., 2007; Atkinson et al., 2008; Dunn et al., 2008; Gray et al., 2009), as well as to address the nature of constraints on linguistic diversity in an evolutionary framework (Dunn et al., 2011).

However, evolutionary studies are not restricted to language evolution. The histories of copied texts, consisting in reproduction and evolution resulting from errors or intentional modifications introduced by copyists, offer yet another system whose dynamics can be naturally investigated with similar mathematical tools. In this context, *textual criticism* is an active research field that is concerned with the identification of textual variants in either manuscripts or printed books, the ultimate objective being the production of a 'critical edition' containing a scholarly curated text. Within this field, *stemmatics* is a rigorous approach to textual criticism introduced by Karl Lachmann, a German philologist and critic, in the 18th century (Grier, 1989). Based on the princi-

ple that 'a community of error implies a unity of origin,' this approach aims at determining the relations among the extant manuscripts so as to place them in a family tree, named *stemma codicum*. Along these lines, the reconstruction of the original version of well-known texts is another challenging and open problem (Platnick and Cameron, 1977; Timpanaro, 1985; O'Hara, 1996; Canettieri et al., 2009). For some famous masterpieces, such as the Divina Commedia of Dante Alighieri, these evaluations are indeed still highly debated (Moore, 1889; Tonello and Trovato, 2013). Despite some initial skepticism on the applicability of phylogenetic methods to the reconstruction of the tree for a set of manuscript copies (Caetlidge, 2001; Hanna, 2000; Jones, 2001), the validity of this approach in this context has been indicated by recent works (Spencer et al., 2004; Bordalejo, 2015; Marmerola et al., 2016).

All the aforementioned applications belong to the class of inverse problems: starting from present, incomplete and often noisy information (DNA or protein data, list of words, corpus of texts), one aims at inferring the most likely evolutionary history that can possibly explain the present observations. In this bottom-up approach, a fundamental issue is the quantitative evaluation of the full inference process. In this respect, the availability of valid benchmarks for determining the reliability of the different methods and algorithms used to reconstruct phylogenetic trees is crucial. A standard way of testing the proposed algorithms is the construction of models to generate artificial phylogenies, so that the algorithmic results can be directly compared with the generated and hence known outcomes of interest (Tria et al., 2010c; Pompei et al., 2010; Desper and Gascuel, 2002). However, in doing that, one makes unavoidable assumptions on the evolutionary processes of interest, which in turn may affect the accuracy of the reconstruction and its evaluation.

In this paper, we present an interdisciplinary approach to face this problem. We introduce Copystree, a web-based game in which users are engaged in the very process of collectively generating, through successive elementary actions of copying, artificial phylogenies of manuscripts. The game is actually meant as an experiment to provide the scientific community with valid benchmarks to test strategies for reconstructing phylogenetic histories. Copystree allows for exhaustive monitoring of all the phases of the emergence of a phylogeny: who did what, at what time, copying from whom, etc. While players enjoy the experience, scholars gain access to an unprecedented set of artificial phylogenies whose evolutionary processes do not obey any predefined theoretical mechanisms, being generated instead with the unpredictability of human ability and creativity.

The paper is organized as follows. In the next section we describe the structure of the game, including details about the strategy adopted to generate artifi-

cial phylogenies, and present the results of the first set of gaming sessions held at Sapienza University of Rome, illustrating the most interesting properties of the dynamical process generated through the game. In Section 3 we describe the distance-based approaches for the phylogenetic reconstruction that can be applied to the analysis of the phylogenies generated with Copystree and, more generally, to the analysis of family trees of copied texts. In the final part of Section 3 we also examine the accuracy of the phylogenetic reconstruction for the phylogenies of the first database we have collected.

2 Copystree

In this section, we present the game/experiment Copystree, giving details about the game and the dynamics leading to an artificial phylogeny of copied texts. A schematic description is presented in Fig. 1. At present, Copystree has only been adopted for specific experimental sessions. Soon it will be released as a web-based game accessible to the general public.

2.1 *The game*

The game is organized in gaming sessions where users (players) are challenged to copy a fragment of a text to the best of their abilities. Each session lasts up to 3 minutes, but users can submit their copy before the session is expired. The text is shown in a non-editable graphic format to avoid cut and paste actions, while users enter their copied text in a standard HTML text field (see Fig. 1). In our experiments, several distinct input texts were initially available as seeds for manuscript phylogenies. We set the length of the texts to be in a range of about 100 words (mean value: 85 words, max value: 111 words, min value: 55 words), so that they could be easily copied in a gaming session while allowing for the emergence of a significant level of variation. The input texts were quite heterogeneous, ranging from fragments of works of modern and classical literature to excerpts of short newspaper articles.

In each session, a player is presented with a fragment within the current phylogeny, i.e. the phylogeny generated until that moment, randomly chosen from a set of fragments available for copy (see below for details about the topology of the emerging phylogeny and the availability of fragments for copying). Once created, a fragment starts aging, mimicking the usual processes of degradation that manuscripts and old books undergo during their lifetime. Each fragment can be copied several times, though each time with a different level of degradation, just like old manuscripts could have been found and copied many times in different periods. A player can play with copies from the same phylogeny for

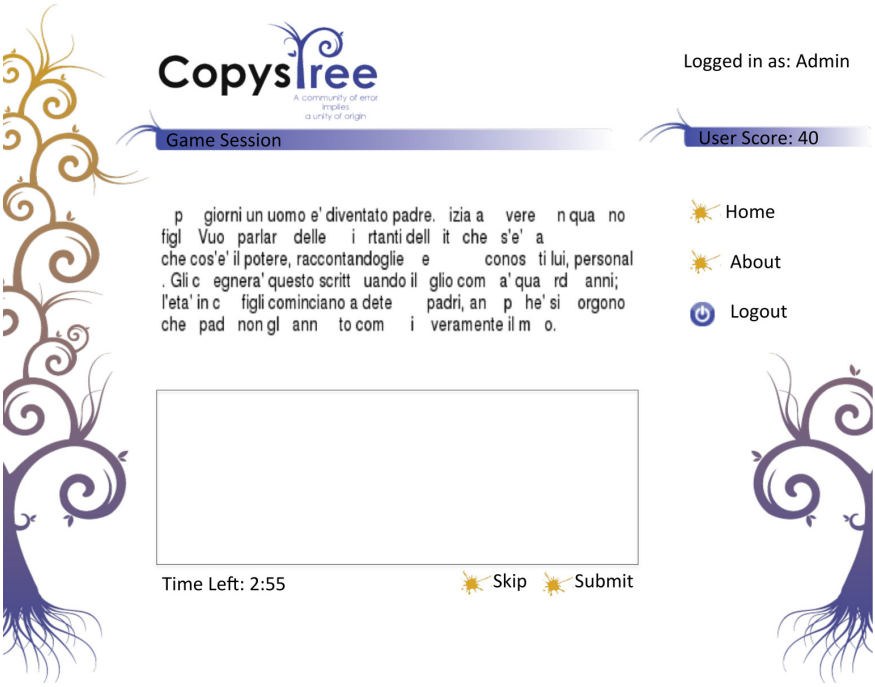

FIGURE 1 *User interface of Copystree. The challenge of the game consists in copying a text to the best of the players' ability, with time constraints and the readability of the text progressively reduced in an artificial way. The text to be copied is presented to the player in a non-editable graphic format, to avoid cut and paste actions, and input is allowed only through a standard* HTML *text field. At the end of each gaming session, players are given a score based on the similarity between the copy they produced and the text they were prompted with. Higher similarities result in higher scores. The scoring system is not explicitly available to players.*

multiple game sessions. In this way we allow for the emergence of horizontal evolution, where the same variant is introduced in two or more independent lineages and will mimic the analogous cases of horizontal gene transfer in biology and borrowings in language evolution. The evolutionary dynamics of each copy is summarized in Fig. 2.

It is important to stress the distinction between a copy of the text and a related artificial text. A copy is the text that a player produces as a result of her/his participation in the game (that is, as a result of her/his copying effort); a related artificial text is a degraded version of a copy, that is, the result of one of the degradation procedures described below. It is this artificial text that is available for a further act of copying by another player (note that many different related artificial texts can be generated from a given copy). In reality, during the actual copying procedure of a manuscript, each variant of the original text

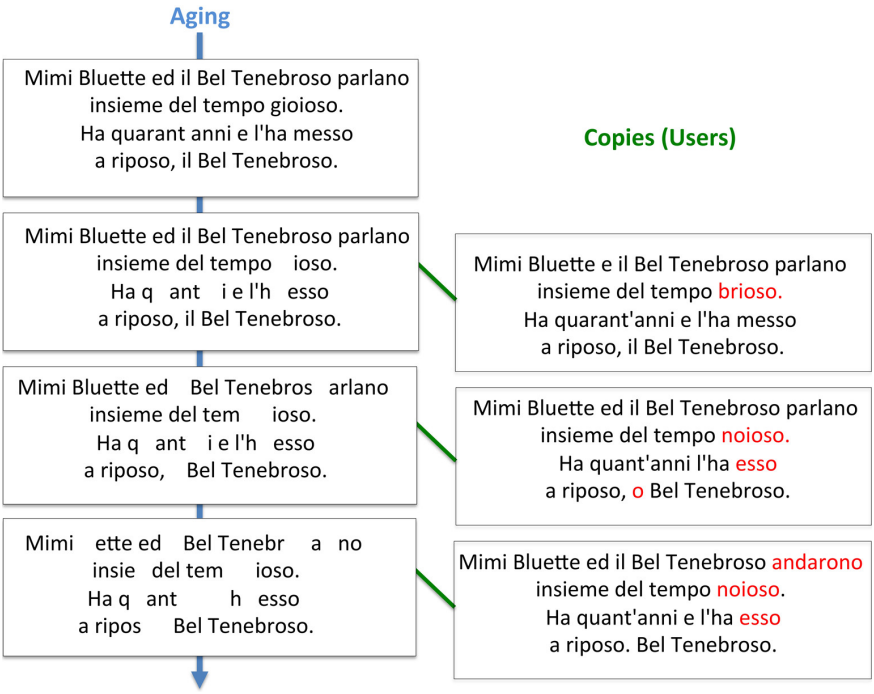

FIGURE 2 *Evolution of a single copy. To mimic the degradation processes that manuscripts and old books undergo during their lifetime, each copy of a text is associated with an independent phylogenetic lineage, through which the text is progressively degraded. Each fragment can thus be copied several times, each time with a different level of degradation, each new copy being the starting point of a new lineage. Because of the reduced readability of the original text, several variants, for example new words (here highlighted in red), may emerge in the new copies.*

(i.e., each copy) was probably copied several times, resulting in a non-binary topology; however, as already discussed above (referring to Fig. 2), in Copystree we constrained each particular artificial text to be copied only once, recovering at this level the binary character of the phylogenetic tree. Each artificial text has two daughter nodes: another artificial text, belonging to the aging lineage and marked with a red square in Fig. 3 A, and a new copy, associated with a new lineage, marked with a green circle in Fig. 3 A. The evolutionary process thus proceeds with a binary structure (Fig. 3 A), and, at the same time, the phylogenetic tree restricted to the copies has a realistic non-binary tree structure (see Fig. 3 B). While the binary inferred trees can easily be compared to non-binary trees (see Section 3 below), a non-binary diversification process could lead to undesired biases for the inference of the correct topology of the tree.

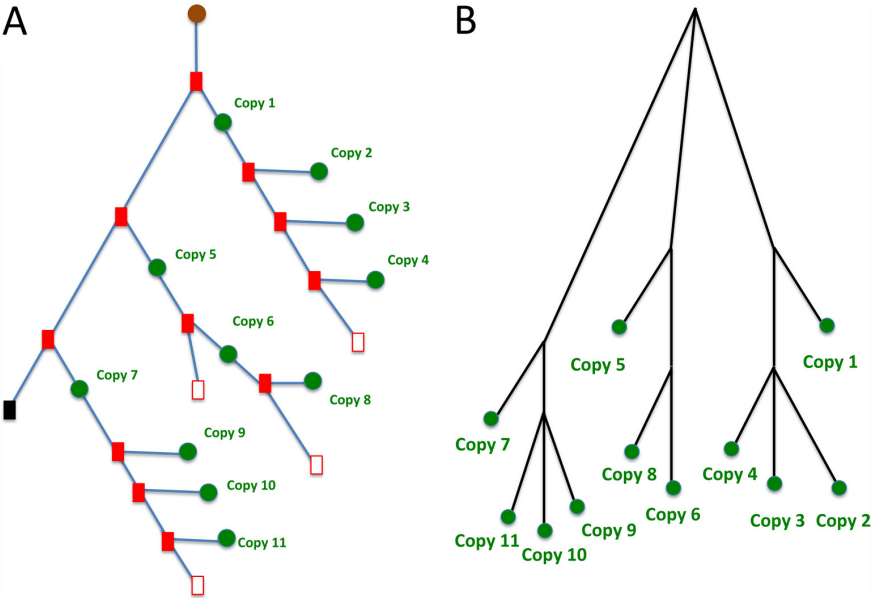

FIGURE 3 *Phylogenetic structures of the artificial phylogenies. A: Schematic illustration of the creation of an artificial phylogeny. Starting from the original text (the root, brown circle), a binary tree is generated via successive copying actions. In each round, a player is presented with a text to copy, chosen from among the elements of the tree available for copying (represented by empty red squares). When copying is completed, the empty square becomes a solid red square and branches into two new nodes of the tree: the copy of the text entered by the player (a green circle) and another empty red square representing the degraded version of the text just copied. This operation is repeated through the successive rounds of the game. At each point in time, the phylogeny consists of a set of artificial texts (squares) and a set of copies (green circles). Only the artificial texts not yet copied (the empty red squares) are available for further copying. Artificial texts are generated to mimic the aging process of each copy, while each copy represents a new phylogenetic lineage (as shown in Fig. 2). Lineages in the tree can be declared inactive and will not be available anymore for copy (black square) if the same fragment is skipped by users more than 3 times. B: A non-binary tree embeds the evolutionary relationship between all the copies of Fig. A. The fact that, in the topology shown in Fig. 3B, the copies 1, 5 and 7 are actually ancestral nodes of the copies below is made explicit by setting the branches above them to have a length = 0. In this way, in the "true phylogeny," which we will use as reference for the inference, all the copies are treated as terminal nodes (this is needed because all inference algorithms will infer a tree where all the copies are leaves), but, on the other hand, we correctly report them as identical to the internal nodes above them.*

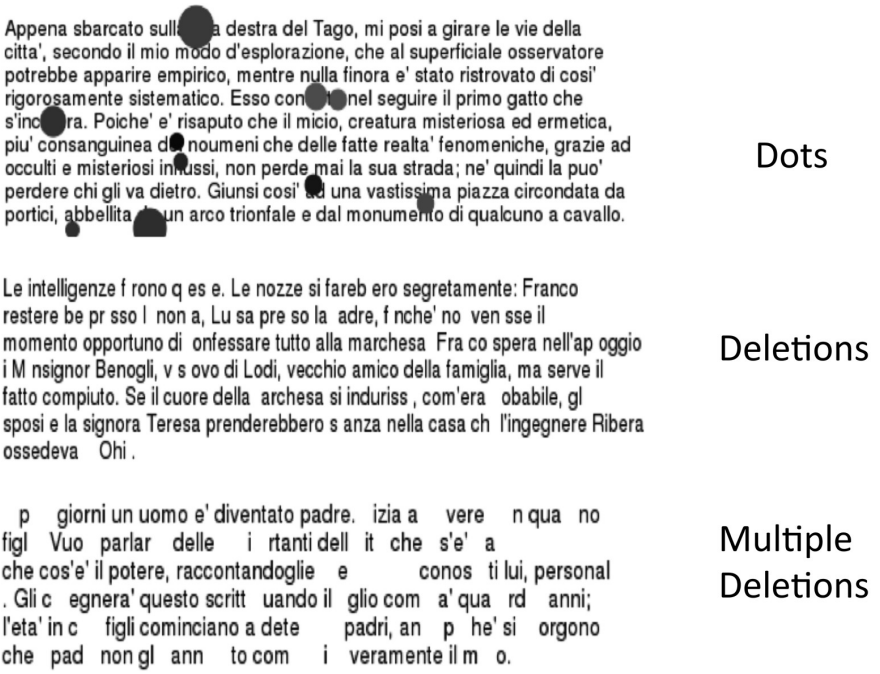

FIGURE 4 *Degradation processes simulating the aging process of a text. Top:* Dots*, where circular colored spots of different sizes are randomly located at different positions to cover portions of the text. Center:* Deletion *of single characters in random positions of the text and replacement with blank spaces. Bottom:* Multiple Deletions*, with the deletion of up to three neighboring characters in randomly chosen locations of the text.*

Let us now describe the different degradation processes implemented in Copystree, through which we simulate the aging process by progressively reducing the readability of each fragment (see Fig. 4). We considered three different degradation processes: the first one uses circular colored spots of different sizes randomly located at different positions to cover portions of the text (*Dots*; Fig. 4, top). In addition, for better control of the *disturbing parameter*, we adopted two further strategies: *Deletion* of single characters in random positions of the texts and replacement with blank spaces (Fig. 4, middle) and *Multiple Deletions*, i.e. deletion of up to three neighboring characters in a randomly chosen location of the text to introduce correlated changes (Fig. 4, bottom). Each degradation process was controlled via a tunable mutation rate, defining the average number of dots and single or multiple deletion events per unit of length of the input text. With the mutation rate we used, the texts were evolving with an average rate of ~ 1.5 degradation units (dots, single characters, groups of 3 contiguous characters) per step. Each of the three strategies mimics the

effects of time on manuscripts and old books, such as paper or ink deterioration, which result in reduced readability and cause increased error rates in the copying procedure. In our experiments, we controlled for the specific degradation method adopted in order to be able to isolate and evaluate separately the effect of the three strategies.

At the end of each game session, the copied text is compared with the text presented to the player, and the similarity between them is computed through the edit (or Levenshtein) distance (see Section 3). Players are given a score based on that similarity. Higher similarities result in higher scores. The cumulative scores of all players are stored and a chart with all the top scores is displayed on the home page of the game. Players can choose between different languages besides their native language. The game performs a quality check of the copied texts to prevent inhomogeneity in the database; copied texts are stored in the database only if the measured similarity with the presented text is higher than a tunable threshold value. Players can also skip any game session and decide to play with another, always randomly chosen, fragment. If the same fragment is skipped more than 3 times, it is declared inactive and will no longer be available for copying (black square in Fig. 3 A). Through this mechanism, lineages in the phylogeny are dynamically selected according to their grammatical and semantical readability.

Each phylogeny is stored with all the information about its evolution: the full topology of the tree, with all the sequences associated with each internal node and the deterioration process used in its evolution; the ID and native language of the player who created the copy; as well as the copying time of the associated game session.

2.2 *Preliminary game session*

To investigate the potential of Copystree, we organized a two-day session of experiments, held at Sapienza University of Rome. During each experiment, lasting about 8 hours, students were invited to play as many rounds of Copystree as they wished, with small prizes (book vouchers) for the first three classified (we considered cumulative points gained in all the gaming rounds). The participation was very heterogeneous, with some students playing only a few rounds and others playing during the whole duration of the experiment (see Fig. 5). During this session, we were able to collect data for several phylogenies of different lengths, with the three different degradation strategies discussed above, namely: (i) circular colored spots, (ii) single character and (iii) multiple character deletions. As the majority of the users were Italian students, most of the phylogenies collected are in Italian, but we also collected a few examples of phylogenies in English and two phylogenies in Latin (see Table 1 and Fig. 5). As

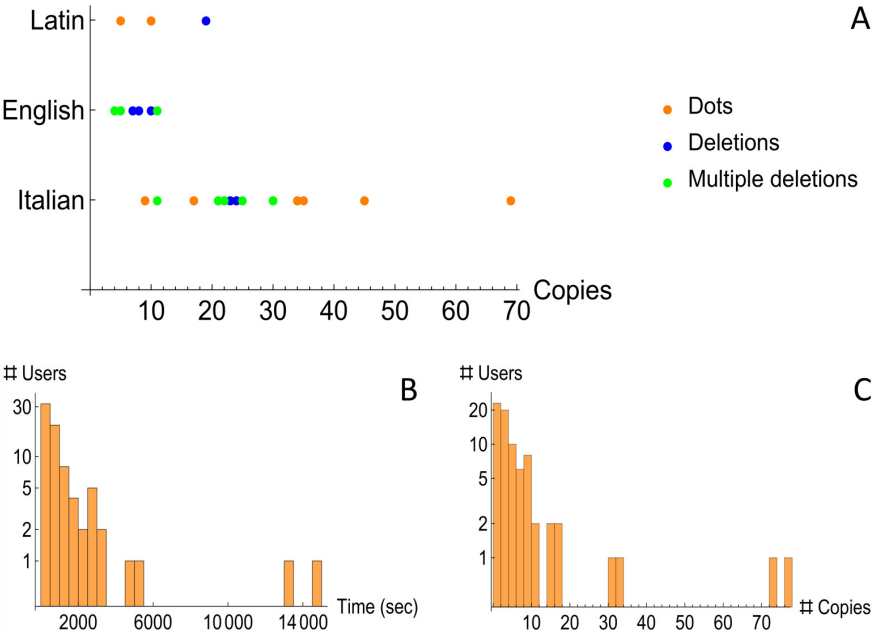

FIGURE 5 *Statistics of the database collected in the preliminary session. A: Scatter plot for the size of the artificial phylogenies (x axis) for the three languages adopted to generate the phylogenies. Different colors denote different degradation processes (see legend). B: Histogram of the cumulative gaming time per user. C: Histogram of the number of copies per user.*

TABLE 1 *Summary statistics of the dataset generated with the first, preliminary session of Copystree. The artificial trees are divided into three classes, corresponding to the three degradation processes considered. (See also Fig. 5.)*

Degradation	Phylogenies	Italian	English	Latin
Dots	9	6	2	1
Deletions	9	6	3	0
Multiple deletions	6	2	3	1

in other similar experiments (Spencer et al., 2004), we decided to include some works of literature in our corpus of texts (see Fig. 6); so players had to deal with unusual uses of language, which would further increase the error rate of the copying procedure.

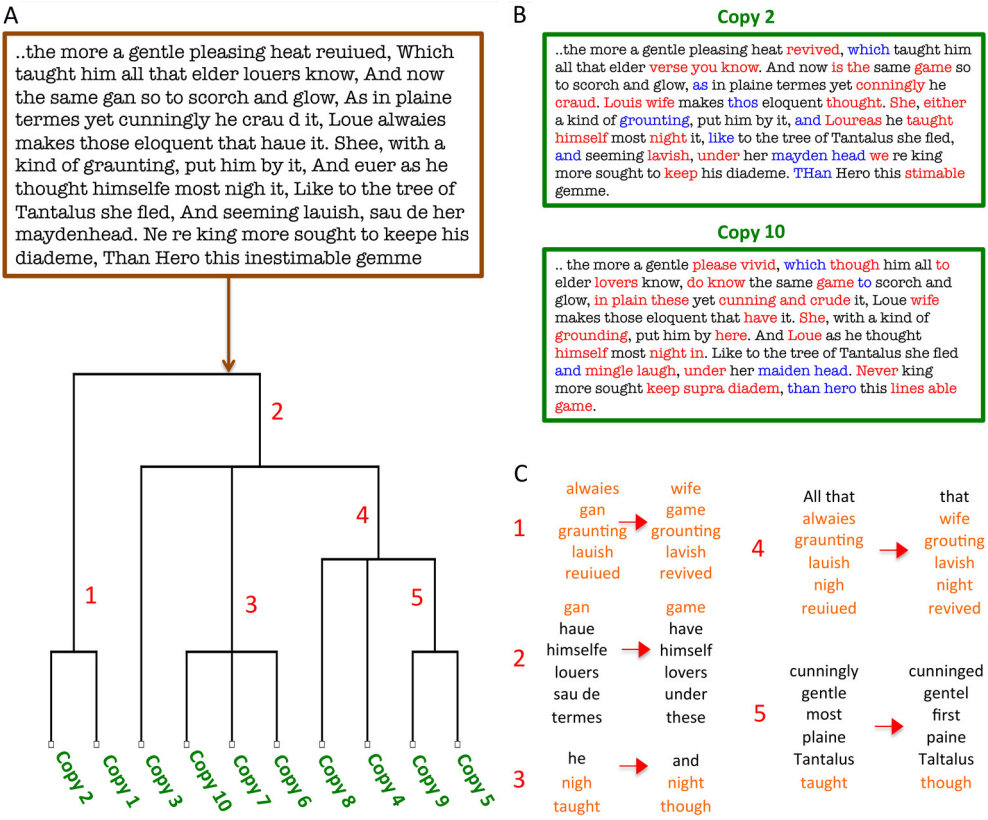

FIGURE 6 *An example of artificial phylogeny. A: The root of the artificial phylogeny, taken from* Hero and Leander *by Christopher Marlowe. During the gaming sessions, this text was copied 10 times, following the scheme illustrated in Fig. 3. Here we report the non-binary phylogeny describing the diversification process of the set of copies. B: Examples of two copies belonging to this artificial phylogeny. The texts differ from the root due to accidental typos (marked in blue) and because new words have emerged during the evolution (marked in red). C: Variants that emerged during the evolution of the text. Numbers indicate the tree branch (as marked in the A panel) where the variant appeared. Several events of parallel evolution can be identified, where the same word has emerged in two independent lineages (words marked in orange).*

2.2.1 An example of a phylogeny created with Copystree

In Fig. 6 we show an example of an artificial phylogeny generated with Copystree. The root text of this phylogeny was extracted from *Hero and Leander* by Christopher Marlowe, written in Early Modern English, and is shown in Fig. 6 A. The phylogeny was generated according to the scheme described in the previous section; in this case, the artificial texts were modified and shown to the

players with the multiple deletions degradation process. Ten different copies were collected, whose evolutionary structure is illustrated in the non-binary tree of Fig. 6 A. As shown in Fig. 6 B, both accidental typos and variants of entire words, resulting in semantic changes, occurred during the evolution of the text. Remarkably, several events of parallel evolution, where the same word emerged in two independent lineages, can be identified.

2.2.2 Complexity of the phylogenies produced by Copystree

The mutational dynamics generated with Copystree display properties that cannot be captured by simple artificial generative algorithms, which are the standard benchmarks for phylogeny reconstruction algorithms (Tria et al., 2010c; Pompei et al., 2010; Desper and Gascuel, 2002). In particular, we find that the amount of variation, quantified with both the edit distance and the mean number of new variants, does not increase linearly with each new copy, as shown in Fig. 7. This effect results from the actions of the users, which introduce changes in the copying procedure that are biased towards the preservation of both the semantics and the correct spelling of words. An example of this process is shown in Fig. 7 C.

An equivalent statement to describe this phenomenon is that the mutational changes between consecutive copies are not independent. This pattern is widely observed in evolutionary biology and is the result of compensatory mutations occurring, for example, when the fitness loss caused by one mutation is remedied by its epistatic interaction with a second mutation at a different site in the genome.

In addition, we found that the observed mutation rate of the copies does not significantly differ between the three degradation strategies, *Dots, Deletion* and *Multiple Deletions* (see Fig. 7), although these were originally introduced in order to tune the disturbing parameter and influence the error rate of the players. This result suggests that the main evolutionary force is associated with the action of the players.

3 Phylogenetic reconstruction

In this section we describe the distance-based approaches for the phylogenetic reconstruction that can be applied to the analysis of family trees of copied texts and, in particular, for the phylogenies generated with Copystree. Other phylogenetic approaches, such as character-based methods (Maximum Likelihood, Maximum Parsimony, Bayesian Analysis), would require a much higher computational cost. Moreover, the evolutionary models that are currently in use

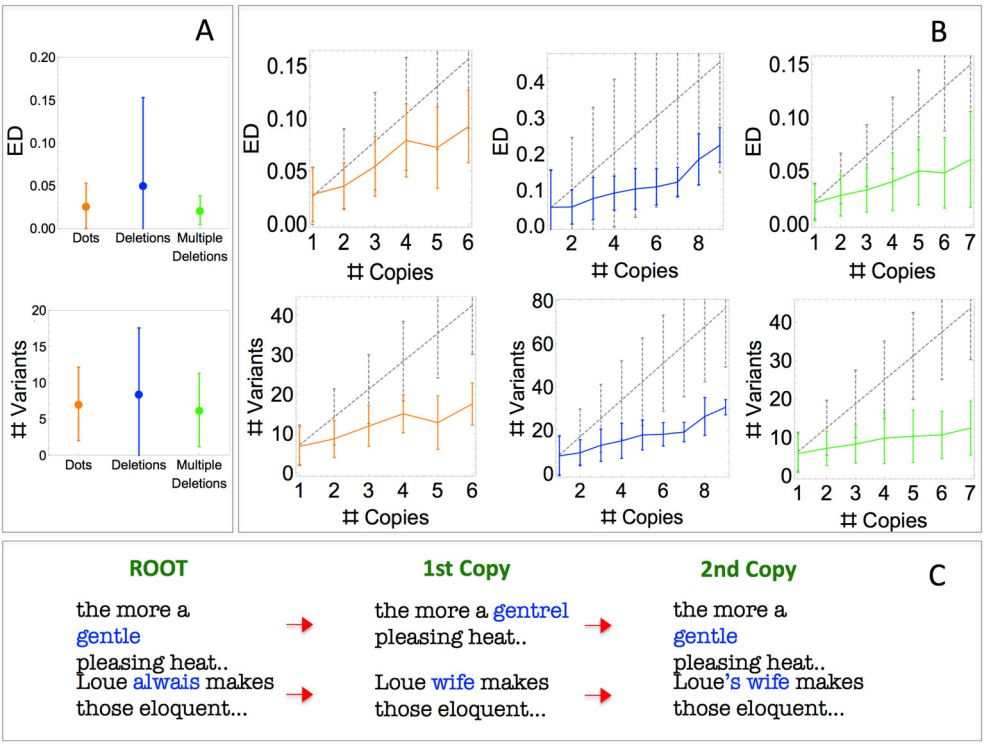

FIGURE 7 *Mutation rates. A: Mean value and standard deviation of the edit distance (top) and number of variants (i.e. different words) measured between two consecutive copies (bottom), for the three different degradation processes considered. B: Same information as in A but evaluated as a function of the number of copies away from the original text. We show in grey the expected value (plus/minus standard deviation) of both the edit distance and the number of observed variants, under the hypothesis of independent changes (i.e. linear extrapolations of the values of A after many copies). C: Examples of the evolution of a text after multiple copies; changes are highlighted in blue. In the first case, a typo introduced after the first copy is restored in the subsequent one. In the second case, the introduction of a new variant in the first copy induces a change of the semantic content of the sentence, which is retained in the subsequent copy. These examples are taken from the tree of copies of* Hero and Leander *by Christopher Marlowe (same as Fig. 6).*

for these approaches were developed in the context of evolutionary biology (see, for example, Drummond and Bouckaert, 2015) and therefore would not be directly applicable to the present research. In the second part of this section, we analyze the accuracy of the phylogenetic reconstruction for the artificial phylogenies of the first database we have collected.

3.1 *Methods*

3.1.1 Alignment, distance and number of variants

The distance between two texts is computed by means of the Levenshtein or edit distance (ED) (Levenshtein, 1966). The ED between two strings is defined as the minimum number of edit operations needed to transform one string into the other, the allowable edit operations being insertion of a character, deletion of a character and substitution of a single character. In addition, the number of aligned words that differ in the two texts also offers a natural measure for the divergence between copied texts. The alignment between copies is performed by means of the Needleman-Wunsch algorithm for sequence alignment (Likic, 2008), with score 1 for matches and score zero for non-matches and gaps.

While the ED, which accounts for a more punctuated description of the differences between texts, is an appropriate measure of the evolutionary distance between copies and can be used for computing the distance matrices used in the inference of the phylogenetic trees, the number of variants offers a more coarse-grained description of the divergence between texts and can be used to detect semantic changes (see Fig. 7c).

3.1.2 Distance-based algorithms

Distance-based phylogenetic reconstruction builds upon the computation of pairwise distances among all the pairs of taxa under consideration. In this context, the definition of distance as well as the properties of the distance matrix represent key parameters. In particular, if the input matrix is additive, i.e., if it can be constructed as the sum of a tree's branch lengths, all the algorithms guarantee the correct reconstruction of the unique true tree. Violations of additivity can arise both from experimental noise and from properties of the evolutionary process underlying the observed data. In particular, two main sources of non-additivity are so-called back-mutations, resulting from multiple mutations in the same character/locus, and horizontal transfer events, where two or more individuals belonging to independent lineages of the tree happen to exchange genetic or linguistic content. For a systematic analysis of the emergence of non-additivity, see, for example, Pompei et al. (2010).

In this study we adopt the standard *Neighbor-Joining* (Saitou and Nei, 1987) algorithm along which its recent, more mathematically founded modification *FastME* (Desper and Gascuel, 2002). We also adopt a Stochastic Local Search algorithm we have recently introduced, named *Fast-SBiX*, which was shown to outperform both *Neighbor-Joining* and *FastME* for the inference of language trees and artificially generated phylogenies (Tria et al., 2010b; Pompei et al., 2010; Tria et al., 2010c; Pompei et al., 2011).

3.1.3 Distances between trees

Two suitable measures for a quantitative comparison between binary phylogenetic trees, as inferred from an algorithm, and non-binary trees have been introduced in Pompei et al. (2011): the Generalized Robinson-Foulds score (GRF) and the Generalized Quartet Distance (GQD), which are generalizations of the Robinson-Foulds (Robinson and Foulds, 1981) and the Quartet Distance measures (Bryant et al., 2000), respectively. The GRF and the GQD offer two complementary quantitative measures of the distance between trees (Christensen et al., 2005; Pompei et al., 2011). The GQD quantifies the number of quartets in the inferred binary tree that are not compatible with the quartets induced by the true, non-binary topology, and is a global measure of the agreement between the two trees, being sensible to the size of misplaced subtrees. The GRF, on the other hand, offers a quantitative assessment on the distance (measured as the number of edges) between subtrees that are moved in one tree with respect to the other.

We here define the two measures in mathematical terms. Let T_e be the topology of a non-binary tree (e.g., the tree of copied texts) and T_i the inferred binary tree, then we define the Generalized Robinson-Foulds (GRF) distance as:

$$GRF(T_i, T_e) = \frac{i(T_i) - e(T_i, T_e)}{i(T_i)} \tag{1}$$

where $i(\mathrm{T}_i)$ denotes the number of internal edges of T_i and $e(\mathrm{T}_i,\mathrm{T}_e)$ the number of bipartitions in T_i compatible with those in T_e. Intuitively, a bipartition in T_i is said to be compatible with a bipartition in T_e if it does not contradict any of the bipartitions induced by cutting an edge in T_e. More rigorously, the compatibility of a bipartition b of T_i with the tree T_e is defined as follows. Let us call b_1 and b_2 the two sets defining b, and a_1^i, a_2^i the two sets defining the i^{th} bipartition of T_e. The partition b is compatible with the tree T_e if for each bipartition i of T_e, the following is true: $b_1 \subseteq a_1^i$, or $b_1 \subseteq a_2^i$, or $b_2 \subseteq a_1^i$, or $b_2 \subseteq a_2^i$ (see Fig. 8). The Generalized Quartet Distance (GQD) is defined as:

$$GQD(T_i, T_e) = \frac{d(T_i, T_e)}{norm(T_e)} \tag{2}$$

where $d(\mathrm{T}_i,\mathrm{T}_e)$ denotes the number of different quartets in T_i and T_e. The normalization factor norm(T_e) is equal to the number of quartets in T_e.

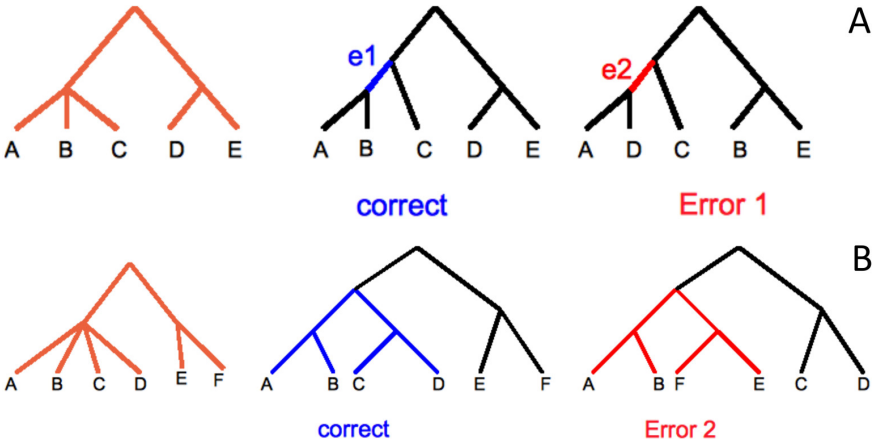

FIGURE 8 *Comparison between binary and non-binary trees. Top: Example of a compatible (blue) and a non-compatible (red) edge between a non-binary tree (left, orange) and a binary tree, as considered in the Generalized Robinson-Foulds distance. Bottom: Example of a compatible (blue) and a non-compatible (red) quartet between a non-binary tree (left, orange) and a binary tree, as considered in the Generalized Quartet Distance.*

3.2 *Accuracy of the reconstruction of phylogenies generated with Copystree*

The database gathered in our experiments already allows for a very first test of the accuracy of phylogenetic reconstruction. The inference framework that can be used in this case is the *distance-based* approach, described in the Methods section (3.1). Phylogenetic trees are reconstructed based on the observed distance matrix, computed among all the taxa present in the dataset. In this context, a proper definition of distance is represented by the edit distance between the aligned versions of the copies (see Section 3.1).

Starting from the distance matrix of each artificial phylogeny, three different algorithms, *Neighbor-Joining* (Saitou and Nei, 1987), *FastME* (Desper and Gascuel, 2002) and *Fast-SBiX* (Tria et al., 2010c; Tria et al., 2010b), have been used for the inference. To quantify the accuracy of the reconstruction, two different measures have been used, namely the Generalized Robinson-Foulds score (GRF) and the Generalized Quartet Distance (GQD), which allow for a quantitative comparison between binary phylogenetic trees, as inferred by one of the three algorithms, and non-binary trees associated with each artificial phylogeny (for details, see Pompei et al., 2011, and Section 3.1).

In Table 2, we report the mean value and the standard deviation of the GQD and GRF computed for all three algorithms and for each class of artificial phylogeny: (i) *Dots*, where the degradation process was produced with colored

TABLE 2 *Accuracy of the reconstruction. Generalized Quartets Distance* (GQD) *and Generalized Robinson Foulds* (GRF) *distance* (*see Section 2*) *between the non-binary tree of copies and the inferred phylogenetic tree, for the distance-based algorithms used for the inference: Fast-SBiX, Neighbor-Joining and FastME. These results can be compared to the mean GRF and GQD values for a set of randomly reconstructed phylogenetic trees* (*column 'random'*), *where, for each tree, we have considered a set of 10 random reconstructions, where the topology is randomly extracted from all the possible trees with the same number of leaves. The artificial trees are divided into three classes, corresponding to the three degradation processes considered.* (*See also Fig. 1.*)

	Fast-SBiX	**FastMe**	**NJ**	**random**
GQD				
Dots	0.19 ± 0.17	0.19 ± 0.17	0.19 ± 0.17	0.52 ± 0.11
Deletions	0.10 ± 0.10	0.15 ± 0.08	0.16 ± 0.09	0.49 ± 0.13
Multiple deletions	0.18 ± 0.20	0.18 ± 0.21	0.18 ± 0.21	0.53 ± 0.07
GRF				
Dots	0.42 ± 0.27	0.42 ± 0.26	0.43 ± 0.25	0.83 ± 0.18
Deletions	0.35 ± 0.24	0.38 ± 0.21	0.30 ± 0.23	0.74 ± 0.19
Multiple deletions	0.50 ± 0.32	0.50 ± 0.36	0.51 ± 0.37	0.85 ± 0.18

dots, (ii) *Deletions*, where texts were degraded with single character deletions, and (iii) *Multiple Deletions*, degradation by correlated deletions. For comparison, we have also measured the GRF and GQD values for a set of randomly reconstructed phylogenetic trees where, for each tree, we have considered a set of 10 random reconstructions, with the topology randomly extracted from all the possible trees with the same number of leaves.

All the inferred phylogenies feature a relatively low value of the GQD, pointing to a general ability to recover the correct topology. However, the GRF values are quite high, showing a significant level of single misplaced taxa (see again Section 3.1). The accuracy of the reconstruction is not significantly affected by the size of the phylogeny (see Fig. S1), and we observe very similar performances among all three distance-based algorithms considered. As the GQD is the best criterion to quantify the overall agreement between the reconstructed tree and the original phylogenies (see discussion in Pompei et al., 2011), we find that the lower mean values of the GQD for the *Fast-SBiX* algorithm indicate a slightly higher average accuracy of the reconstruction for this algorithm, which is consistent with a previous analysis (Pompei et al., 2011).

4 Conclusions

Phylogenetic reconstruction is a common framework for the analysis of evolutionary processes in several research fields. Distance-based approaches, in particular, offer a flexible mathematical tool, since the only requirement is a dissimilarity matrix, computed by means of a suitable distance between pairs of taxa. This class of algorithms, featuring a very low computational complexity, is particularly suitable for tackling phylogenetic reconstructions of large datasets.

While all distance-based algorithms correctly infer the unique phylogenetic tree associated with an additive distance matrix, i.e., a dissimilarity matrix where all the pairwise distances can be expressed as the sum of branch lengths of a tree connecting all the taxa, violations of the additivity condition typically occur when it comes to distances observed in both biological and linguistic data. As in many other inverse problems, the main source of benchmarks for assessing the ability of the different algorithms are artificially generated phylogenies, often produced through simple evolutionary processes, where hypotheses on the underlying evolutionary mechanisms are unavoidable. This procedure presents an intrinsic limitation: when dealing with real datasets, one typically does not know which model of evolution is the most suitable for them.

Here we have presented a web-based game that offers the unprecedented opportunity to generate artificial phylogenies in a highly monitored and controllable setup. The idea behind Copystree is to mimic the evolution of manuscripts resulting from errors or intentional modifications that occur during copying by human players. In Copystree, all the essential information is available since the game records every single detail of the gaming sessions. Further, the evolution of the manuscripts is not driven by any forces determined a priori, but, instead, it is the outcome of a collective copying process.

We presented the results of a first set of experiments where Copystree was deployed to generate artificial phylogenies. It turns out that the resulting phylogenies feature a realistic level of complexity that is hardly observed with simple generative algorithms. For instance, we find parallel evolution, where the same variant of a word emerges in two independent lineages, and compensatory changes, which result from the abilities and decisions of players who, while copying, try to restore the semantic and grammatical correctness of the copied text. Though limited in size, the gathered datasets already allow a first comparison of the accuracy of several phylogenetic reconstruction algorithms. We compared three distance-based algorithms, *Neighbor-Joining* (Saitou and Nei, 1987), *FastME* (Desper and Gascuel, 2002) and *Fast-SBiX* (Tria et al., 2010c; Tria et al., 2010b).

Two main findings emerged. Firstly, the three algorithmic schemes performed very similarly (except for a slightly higher average accuracy of *Fast-SBiX*, a result in accordance with previous investigations; cf. Pompei et al., 2011). Secondly, and more importantly, the performances of the three phylogenetic algorithms considered here were far from being without error. This implies that the evolutionary dynamics developed by Copystree represent a real challenge for most of the current phylogenetic tools. Thus, a word of caution is in place when it comes to applying current phylogenetic tools and measuring their performance based on artificially generated benchmarks.

Perhaps it is worth rethinking the way in which we assess the accuracy of phylogenetic algorithms. From this perspective, Copystree could represent an important stepping stone, which might have an impact in all active research fields in which phylogenetic classifications are relevant—ranging from evolutionary biology (Simonson et al., 2005) to immunology (Holmes and Grenfell, 2009; Grenfell et al., 2004) and historical linguistics (Renfrew et al., 2000; Joseph and Janda, 2004; Wichmann and Grant, 2012; Gray and Atkinson, 2003; Bryant et al., 2005; Pagel et al., 2007; Atkinson et al., 2008; Dunn et al., 2008; Gray et al., 2009; Holman et al., 2011; Jäger, 2013, 2015; Holman and Wichmann, 2015; Jäger, 2014).

Supplementary material

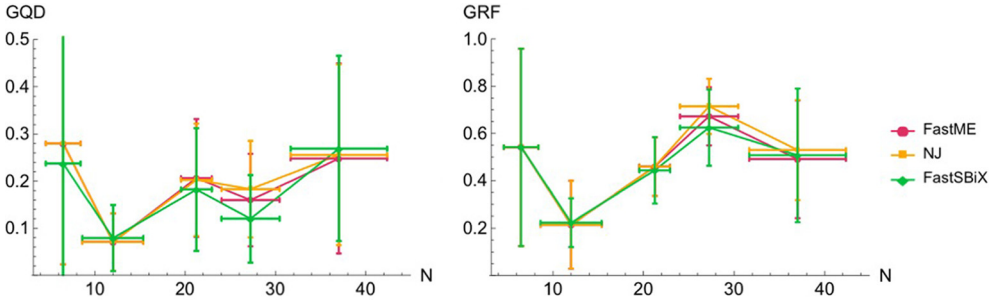

FIGURE S1 *Accuracy of the reconstruction. We study here the accuracy of the reconstructed phylogenetic trees of our dataset as a function of the size of the phylogeny, i.e., the amount of copied text. In this plot we include all the phylogenies of our dataset (i.e., all three degradation processes considered together); trees are then grouped into classes of 5 elements, for which we show the mean value of the* GQD *and* GRF *(y axis) as a function of the mean size of the phylogeny* N *(x axis). Trees were reconstructed with the three distance-based algorithms considered in this context (see main text):* FastME, Neighbor-Joining (NJ) *and* Fast-SBiX.

References

Atkinson, Quentin D., Andrew Meade, Chris Venditti, Simon J. Greenhill, and Mark Pagel. 2008. Languages evolve in punctuational bursts. *Science* 319(5863): 588–588.

Bordalejo, Barbara. 2015. The genealogy of texts: Manuscript traditions and textual traditions. *Digital Scholarship in the Humanities* 31(3): 563–577.

Bryant, David, Flavia Filimon, and Russell D. Gray. 2005. Untangling our past: Languages, trees, splits and networks. In Ruth Mace, Clare J. Holden, and Stephen Shennan (eds.), *The Evolution of Cultural Diversity: A Phylogenetic Approach*, 67–83. Walnut Creek, CA: Left Coast Press.

Bryant, David, John Tsang, Paul E. Kearney, and Ming Li. 2000. Computing the quartet distance between evolutionary trees. In David Shmoys (ed.), *Symposium on Discrete Algorithms: Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, 285–286. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Caetlidge, Neil. 2001. The Canterbury Tales and cladistics. *Neuphilologische Mitteilungen* 102(2): 135–150.

Canettieri, Paolo, Vittorio Loreto, Marta Rovetta, and Giovanna Santini. 2009. Philology and information theory. *Cognitive Philology* 1: 1. Downloadable at http://ojs.uniroma1.it/index.php/cogphil/article/view/8816/8797 (accessed February 20, 2018).

Chris Christiansen, Thomas Mailund, Christian N.S. Pedersen, and Martin Randers. 2005. Computing the quartet distance between trees of arbitrary degree. In Rita Casadio and Gene Myers (eds.), *Algorithms in Bioinformatics. 5th International Workshop, WABI 2005*, Lecture Notes in Bioinformatics 3692, 77–88. Berlin: Springer.

Darwin, Charles R. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Darwin, Charles. 1871. *The Descent of Man, and Selection in Relation to Sex*. London: John Murray.

Desper, Richard and Olivier Gascuel. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology* 9(5): 687–705.

Drummond, Alexei J. and Remco R. Bouckaert. 2015. *Bayesian Evolutionary Analysis with Beast*. Cambridge: Cambridge University Press.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473(7345): 79–82.

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink, and Angela Terrill. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84(4): 710–759.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

Gascuel, Olivier. 2005. *Mathematics of Evolution and Phylogeny*. Oxford: Oxford University Press.

Gray, Russell D. and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965): 435–439.

Gray, Russell D., Alexej J. Drummond, and Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science* 323(5913): 479–483.

Grenfell, Bryan T., Oliver G. Pybus, Julia R. Gog, James L.N. Wood, Janet Daly, Jenny A. Mumford, and Edward C. Holmes. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303(5656): 327–332.

Grier, James. 1989. Lachmann, bédier and the bipartite stemma: Towards a responsible application of the common-error method. *Revue d'histoire des textes* 18(1988): 263–278.

Hanna, Ralph. 2000. The application of thought to textual criticism in all modes—with apologies to A.E. Housman. *Studies in Bibliography* 53: 163–172.

Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, and others. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology* 52(6): 841–875.

Holman, Eric W. and Søren Wichmann. 2017. New evidence from linguistic phylogenetics supports phyletic gradualism. *Systematic Biology* 66.4: 604–610.

Holmes, Edward C. and Bryan T. Grenfell. 2009. Discovering the phylodynamics of RNA viruses. *PLoS Computational Biology* 5(10): e1000505.

Jäger, Gerhard. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* 3(2): 245–291.

Jäger, Gerhard. 2014. Evaluating distance-based pyhlogenetic algorithms for automated language classification. Technical report, University of Tübingen. Downloadable at http://www.sfs.uni-tuebingen.de/~gjaeger/publications/njFastme.pdf (accessed February 20, 2018).

Jäger, Gerhard. 2015. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences of the U.S.A.* 112(41): 12752–12757.

Jones, Alex. 2001. The properties of a stemma: Relating the manuscripts in two texts from the Canterbury Tales. *Parergon* 18(2): 35–53.

Joseph, Brian D. and Richard D. Janda (eds.). 2004. *The Handbook of Historical Linguistics*. Malden, MA: Blackwell Publishing.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady* 10: 707–710.

Likic, Vladimir. 2008. The Needleman-Wunsch algorithm for sequence alignment. Lecture given at the 7th Melbourne Bioinformatics Course of the Bi021 Molecular Science and Biotechnology Institute, University of Melbourne. Lecture notes downloadable at https://www.cs.sjsu.edu/~aid/cs152/NeedlemanWunsch.pdf (accessed February 20, 2018).

Marmerola, Guilherme D., Marina A. Oikawa, Zanoni Dias, Siome Goldenstein, and Anderson Rocha. 2016. On the reconstruction of text phylogeny trees: Evaluation and analysis of textual relationships. *PloS One* 11(12): e0167822.

Maynard Smith, John and Eörs Szathmáry. 1997. *The Major Transitions in Evolution*. Oxford: Oxford University Press.

Moore, Edward. 1889. *Contributions to the Textual Criticism of the Divina Commedia*. Cambridge: Cambridge University Press.

O'Hara, Robert J. 1996. Trees of history in systematics and philology. *Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano* 27: 81–88.

Pagel, Mark, Quentin D. Atkinson, and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163): 717–720.

Platnick, Norman I. and H. Don Cameron. 1977. Cladistic methods in textual, linguistic, and phylogenetic analysis. *Systematic Biology* 26(4): 380–385.

Pompei, Simone, Emanuele Caglioti, Vittorio Loreto, and Francesca Tria. 2010. Distance-based phylogenetic algorithms: New insights and applications. *Mathematical Models and Methods in Applied Sciences* 20(suppo1): 1511–1532.

Pompei, Simone, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS One* 6(6): e20109.

Renfrew, Colin, April McMahon, and Robert Lawrence Trask. 2000. *Time Depth in Historical Linguistics*. Cambridge: The Macdonald Institute for Archaeological Research.

Robinson, David F. and Leslie R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53(1–2): 131–147.

Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4): 406–425.

Simonson, Anne B., Jacqueline A. Servin, Ryan G. Skophammer, Craig W. Herbold, Maria C. Rivera, and James A. Lake. 2005. Decoding the genomic tree of life. *Proceedings of the National Academy of Sciences of the U.S.A.* 102(suppl 1): 6608–6613.

Spencer, Matthew, Elizabeth A. Davidson, Adrian C. Barbrook, and Christopher J. Howe. 2004. Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology* 227(4): 503–511.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96(4): 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21(2): 121–137.

Timpanaro, Sebastiano. 1985. *La genesi del metodo del lachmann* (Vol. 5). Torino: Liviana.

Tonello, Elisabetta and Paolo Trovato. 2013. *Nuove prospettive sulla tradizione della "commedia": seconda serie (2008–2013)*. Limena: libreriauniversitaria.it.

Tria, Francesca, Emanuele Caglioti, Vittorio Loreto, and Andrea Pagnani. 2010a. A stochastic local search approach to language tree reconstruction. *Diachronica* 27(2): 341–358.

Tria, Francesca, Emanuele Caglioti, Vittorio Loreto, and Andrea Pagnani. 2010b. A stochastic local search algorithm for distance-based phylogeny reconstruction. *Molecular Biology and Evolution* 27(11): 2587–2595.

Tria, Francesca, Emanuele Caglioti, Vittorio Loreto, and Simone Pompei. 2010c. A fast noise reduction driven distance-based phylogenetic algorithm. In Hamid R. Arabnia, Quoc-Nam Tran, Rui Chang, Matthew He, Andy Marsh, Ashu M.G. Solo, and Jack Y. Yang (eds.), *Proceedings of BIOCOMP 2010*, 375–380. Athens, GA: CSREA Press.

Wichmann, Søren and Anthony P. Grant. 2012. *Quantitative Approaches to Linguistic Diversity: Commemorating the Centenary of the Birth of Morris Swadesh*. Amsterdam: John Benjamins Publishing.