Latest preliminary version of my habilitation thesis. The final version, including an acknowledgement section and possibly other improvements after the defense, will be available on HAL.

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Paris-Est Sup
Laboratoire d'Informatique Gaspard-Monge (UMR 8049)

Proximity, Similarity and Heredity: From Bioinformatics to Digital Humanities

Presented by Philippe Gambette

Habilitation in Computer Science

Version of October 23, 2024

Olivier Kraif	Professor, Université Grenoble Alpes	Reviewer
Gregory Kucherov	Research Director, CNRS	Examiner
Glenn Roe	Professor, Sorbonne Université	Examiner
Marie-France Sagot	Research Director, Inria	Reviewer
Katherine St. John	Professor, City University of New York	Reviewer
Hélène Touzet	Research Director, CNRS	Examiner

Contents

Introduction

Qui me consolera ? — « Moi seule, a dit l'étude ;
« J'ai des secrets nombreux pour ranimer tes jours. » —

Marceline Desbordes-Valmore, « Le Secret perdu »,
Poésies inédites, Genève, Fick, 1860.

Who will console me? "I," said Study. "See
"Mine endless secrets that shall make thee glad."

Yvor Winters, "The Lost Secret, from the nineteenth century French",
The Collected Poems of Yvor Winters, Manchester, Carcanet, 1978.

From bioinformatics to digital humanities

This document presents my research work conducted during my postdoc at Aix-Marseille Université in 2010/2011 and since September 2011 at Université Paris-Est Marne-la-Vallée, which became Université Gustave Eiffel in 2020, in the Laboratoire d'Informatique Gaspard-Monge.

Whereas the research done during my doctorate at Université Montpellier 2 was mostly focused on the design of combinatorial methods for bioinformatics, and more specifically phylogeny, I started a few collaborations on computer-assisted textual analysis. This part of my work has been gradually increasing, and for a few years I have tried to be active in both fields, with a priority for the bioinformatic study of phylogenetic networks, approached from the point of view of a computer scientist.

The *délégation CNRS* that I was granted in 2020 to have a full semester of research at Lattice (a laboratory in the south of Paris, specialized on linguistics and natural language processing), allowed me to transition more clearly towards digital humanities, and more precisely towards the development of several tools and methods to analyze collections of texts. In this evolution of my research field, I was lucky to have the support of my research unit, LIGM. Marie-Pierre Béal, Cyril Nicaud and Stéphane Vialette have always been supportive, even if this new field was not exactly present in the laboratory. Interactions with colleagues in the bioinformatics and computational linguistics teams of LIGM who have been involved in research projects or in collaborations on articles with me on these topics, were also very valuable.

This transition was also encouraged by positive experiences in the first two hackathons of the Bibliothèque nationale de France, the French national library, in 2016 and 2017. I was a member of the winning team for both events, which consisted in developing digital tools using resources from the national library. In 2016, I was a member of the *Gallicarte* team which created an interface to display on a map the results of a query to the digital library Gallica[1]. In 2017, MusiViz consisted in providing enriched visualizations of audio documents in Gallica, both with a spectrogram of the audio signal and with other text or multimedia documents from Gallica.

The field of my research topics also expanded thanks to natural links between the two fields of bioinformatics and digital humanities, where sequences of text, as well as trees and graphs, play a major role (Gus97; Val02; Mor05). Both require interdisciplinary work, with an effort to understand the needs and constraints of people who do not have a computer science background. In both cases, I also try to identify problems which are relevant for a theoretical analysis in computer science, combinatorially or algorithmically, even if this results in oversimplifying the problem faced in biology or in literature.

I like this approach for two reasons. First, I think simple models of complex phenomenons can provide an easy way to explain the chosen approach to people who are not specialists in algorithmics, and be used as a starting point (possibly with an implemented prototype) to discuss results with experts from the other field. This approach allows to understand together, with these experts, precisely what needs to be improved in future, more complex versions of the developed application. It also avoids working on complex solutions which do not actually solve the problem, because of misunderstandings or unrealistic hopes about the outcome of the collaboration.

Furthermore, focusing on simple models also allows to use classical results from several fields of computer science, such as graph theory or stringology, or to get inspired by simple principles in other applied fields, such as reusing a bioinformatic approach studying the evolution of species to study the history of manuscripts. I think it is useful to wonder if the core of the approach in one field really applies for another application, and whether it just needs to be adapted for the new context, or whether a totally new approach has to be adopted.

So far, I feel that this approach of interdisciplinary work associated with a continued emphasis on theoretical computer science and algorithmics has been quite productive, and this is the way I want to continue developing my research. I have the impression that while several colleagues have a similar approach in bioinformatics, I have had fewer opportunities to observe it in digital humanities, where some of the works presented in this document have convinced me that it can also be fruitful.

On proximity, similarity and heredity

The research project I designed for future research during my postdoc focused on the exploration of neighborhoods in graphs, with applications in bioinformatics and natural langage

[1] The prototype which was developed during the week-end of the event is available at https://igm.univ-mlv.fr/~gambette/gallicarte/. The project was then extended by the Bibliothèque nationale de France, with a collaborative web interface to add geographical coordinates to documents in Gallica, see https://gallica.bnf.fr/blog/21032018/gallicarte-arrive-dans-gallica.

processing. Looking back at the past 13 years, I see that what actually interested me in the notion of neighborhood was how it was often combining the ideas of proximity and similarity, for the elements of the neighborhood.

But while focusing on neighborhoods, I overlooked the relationship between similarity and heredity which was my first motivation to do research in phylogenetics. This field of research aims at reconstructing the evolution of species. The first phyogenetic tree reconstruction algorithm I had seen, UPGMA, consists in deducing heredity information from similarity information between the species.

And at the time, I also did not expect that my research would also lead me to develop links between similarity and heredity in completely different contexts, such as genetic editing of texts or modernisation of old texts in French. I did not expect either that I would explore some links between proximity and heredity through works on cultural or industrial heritage.

The idea of finding similarities has always been very attractive to me in mathematics and computer science. Finding a nice bijection between mathematical objects, building a reduction from one problem to another in order to solve the first one, or to prove NP-completeness of the second one, are tasks I enjoy in my own research and results I like to discover in the works of others. But from a more general perspective, I also like to unveil analogies. This can be quite productive in adapting results from one field to another, and this manuscript will illustrate several contexts where such analogies helped me to adapt methods from bioinformatics to digital humanities in the last decade.

Similarly, I have realized how proximity plays an important role in my research practice. Spending a significant amount of time with colleagues – either in the context of a one-week visit in an international research collaboration, or in the context of several regular meetings with interns, doctoral candidates or colleagues from another field – is a good way not only to get research results but also to see other people's practices, which is especially important for interdiscplinary research. For several research results I obtained, I can acknowledge specific collaborative sessions which were instrumental for the outcome of my research. I therefore chose to include in this manuscript not only the research outcome but also to give some information on the context where it was obtained. I will argue that this context is important for the kind of research I am now planning to make and to supervise in digital humanities, where I try not only to develop new algorithms, sometimes based on a nice mathematical model, but also to provide colleagues of other fields with useful tools. I think that both perspectives are nourished by a wide knowledge of the data or problems we work on, which requires a close proximity with experts in the humanities.

Finally, academic heredity also plays an important role in the way I have developed my research in the last decade. I feel indebted to the supervisors or mentors I was lucky to work with as an early-stage researcher, Olivier Gascuel and Denis Bertrand, Daniel Huson, Michel Habib, Vincent Berry and Christophe Paul, Katharina Huber, Alain Guénoche and Jean Véronis. I still share some of the lessons I learned from them with interns or doctoral candidates I supervise.

My supervision experience also benefited from the inheritance shared by experienced members of two associations, CJC (Confederation of early-stage researchers) and ANDès

(Association nationale des docteurs), where Sylvain Collonge, Alban Cornillet, Florent Olivier, Cécile Frolet, Simon Thierry, Marie-Ange Ventura, Jean-Tristan Brandenburg, Philippe Gauron, Juliette Guérin and Clément Courvoisier among others, helped to get me a broader perspective on research and higher education. The last decade has also given me the oportunity to get a particular attention to gender approaches in research and ways of promoting gender equality in research and higher education. I feel especially grateful to Anna-Livia Morand, Carole Chapin, Jeanne Chiron, Caroline Trotot, Claire Hancock, Nicole Dufournaud and Christine Planté for the knowledge I inherited from them on these topics.

About the content of this document

This document is not a book about a scientific concept or issue, neither a "thèse d'État[2]". I consider this "synthèse de mon activité scientifique[3]" a unique opportunity to describe my scientific journey from bioinformatics to digital humanities, and the beginning of my experience as a research supervisor. Therefore, this will not only be a retrospective presentation of the work done since the end of my doctorate and highlighting the concepts which played a major role in them, similarity, proximity and heredity, but also an overview of the many different ways I am conducting research.

Therefore, after structuring the contents into topics with some kind of manual agglomerative clustering procedure, I tried to illustrate the dynamics of this research. To this aim, I describe the obtained scientific results and my contribution, but also what I consider to be key elements to have a better understanding of the context, either about previous research by myself or others on the topic, about decisive contribution by coauthors or about follow-ups, possibly obtained by other teams. I hope that this will result in a more accurate description of my research work, impacted by science policies at different levels as well as the oportunities and constraints of scientific collaborations.

However, it should also be clear that this document does not fully include the research work I have done after my doctorate, either because it was not finalized or because it did not fit with the notions of similarity, proximity and heredity I want to focus on here.

[2] The "doctorat d'État" was replaced in 1984 by the "habilitation à diriger des recherches" in France.

[3] This definition provided by French regulation, more precisely by the *Arrêté du 23 novembre 1988 relatif à l'habilitation à diriger des recherches*, could be translated as "a summary of my scientific activity".

CHAPTER 1

INFERRING BIOLOGICAL HEREDITY FROM SIMILARITY: RECONSTRUCTING AND ANALYZING PHYLOGENETIC NETWORKS

Phylogenetic networks are networks representing the evolution of species in cases where we do not want to only represent *vertical* transfers of genetic material, from an ancestral species to the species which evolved from it, but also *horizontal* transfers of genetic material between species.

We then use the mathematical model of a *graph* connecting some *nodes* representing current species with other connected nodes, either representing ancestral species, in the case of *explicit* phylogenetic networks, or helping to give an abstract overview of the complexity of horizontal transfers, in the case of *abstract* phylogenetic networks. The connections, called *edges* if they are not directed, and called *arcs* if we want to direct them, usually from the parent species to the child species, represent biological heredity.

Several approaches have therefore been developed in the last 40 years to reconstruct phylogenetic networks from several kinds of data. Following my doctoral thesis on combinatorial methods for phylogenetic network reconstruction (Gam10), I have investigated several mathematical and structural aspects of phylogenetic networks in order to provide a better understanding of these objects and to obtain efficient algorithms for their reconstruction. I have also made available several tools for the phylogenetic network community. I have focused mainly on explicit phylogenetic networks where internal nodes represent extinct species, but some of my works also focused on their links with abstract networks, which I started to develop while writing my doctoral thesis in 2010 (which were published in 2012 in an article with my doctoral supervisors, Vincent Berry and Christophe Paul (GBP12)).

More formally, explicit phylogenetic networks are usually defined as directed *acyclic* graphs, that is networks where it is not possible to find a sequence of nodes, starting and ending by the same node, such that consecutive nodes are connected by an arc. An arc connects a *parent* to a *child*, with the convention that all individuals in the species represented by the child were descendants of individuals represented by the parent. We often consider *binary* phylogenetic networks, containing only, as illustrated in figure 1.1:

- *tree nodes* of *outdegree* 2, that is having two children, and *indegree* either 1, that is having one parent, or 0, that is having no parent, and therefore being called the *root* of the phylogenetic network ; such tree nodes represent a *speciation* event, when a species is split in two;
- *hybrid nodes* with indegree two and outdegree one; such hybrid nodes represent a *reticulation* event which could be for example a horizontal gene transfer, a hybridization or an endosymbiosis;
- *leaves* of indegree 1 and outdegree 0, which are labeled by distinct labels, which represent current species.

These mathematical objects generalize *phylogenetic trees*, which are a special case of phylogenetic networks with no hybrid nodes. We can define more general explicit phylogenetic networks which may not be binary, in this case tree nodes may have outdegree larger than 2 and hybrid nodes, indegree or outerdegree larger than 2 and 1 respectively.

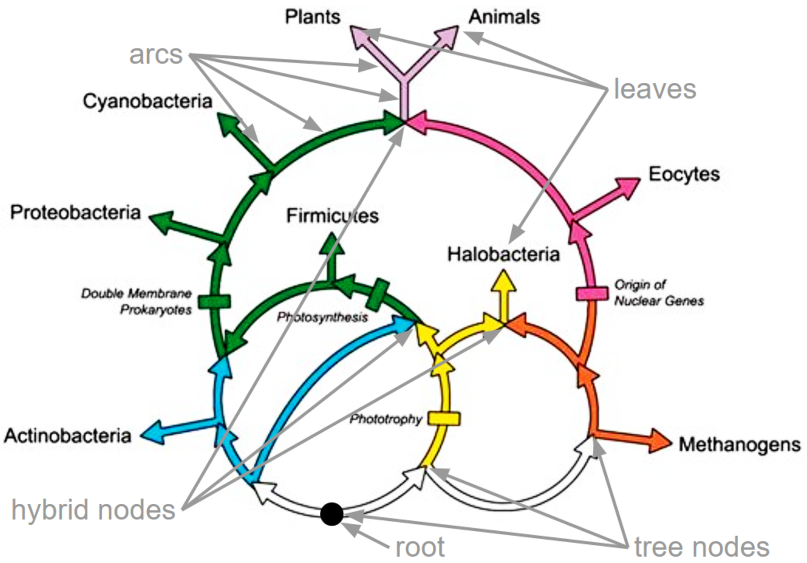


Figure 1.1: A phylogenetic network representing several groups of species, derived from figure 1 of (LLS+15).

1.1 Providing tools for phylogenetic network reconstruction approaches

1.1.1 Keeping up-to-date with research on phylogenetic networks

Until 2020, I continued to develop the website *Who is who in Phylogenetic Networks*, a bibliographic database about methodological papers dealing with phylogenetic networks, which I built at the beginning of my doctorate. The website, available at https://phylnet.univ-mlv.fr, was a useful tool not only for me, to have a good overview of the research work on phylogenetic networks, but also for the phylogenetic network community. I supervised an intern from IIT Ropar in India, Tushar Agarwal, in Summer 2016, to bring some improvements to the website. In particular, four distinct use cases of the website were highlighted on the front page: find experts, explore research, discover software and follow the community. Tushar also coded various kinds of diagrams, to show for example the evolution of the number of different types of publications (see figure 1.2), or of the number of publications for each keyword on the website, as well as co-author graphs with the possibility of highlighting the people working on some topics. In October 2016, this resulted in a report written also with the collaboration of David Morrison, who had already published a survey on practical uses of phylogenetic network tools and methods (AGM16).

In this report, we also analyzed some of the data gathered on the website. For example, using the factor analysis tool of Lexico 3 (LMF+02), which allows to visualize a projection of textual documents in a 2-dimension space based on their lexical similarities, we studied the abstracts of publications gathered in the website. We observed that the main axis of the visualisation puts publications with the most theoretical content on one side, in mathematical or computer science, and publications with the most practical content, with links to biology,

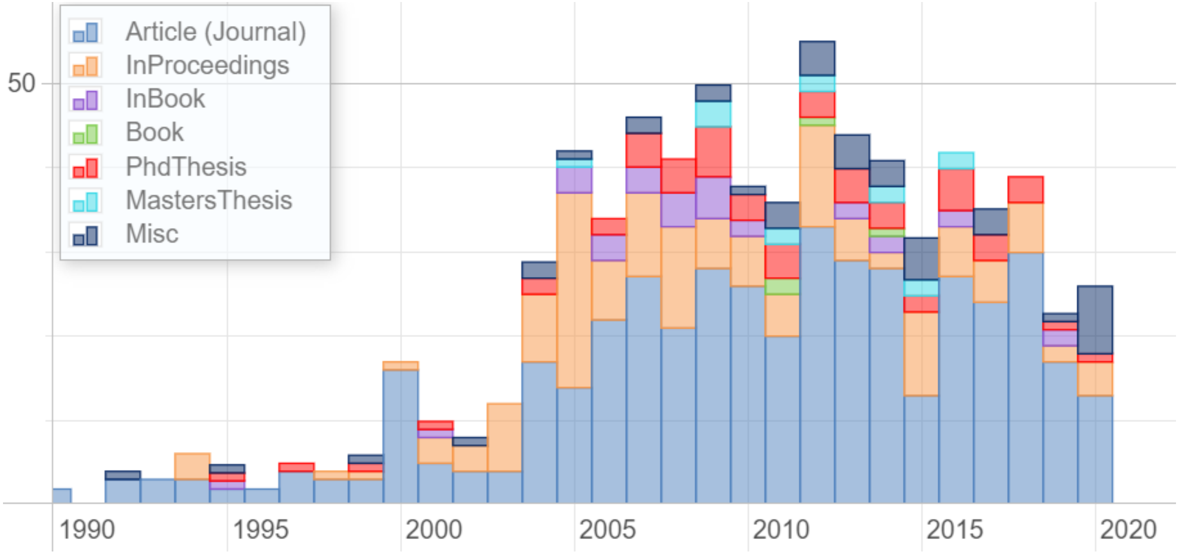


Figure 1.2: Diagram of the 752 publications gathered in the website *Who is who in phylogenetic networks* by type and publication year.

on the other side. This allows to classify publications on the spectrum of bioinformatics, from biology to theoretical computer science.

Other improvements were added to the front page of the website during the workshop on *Distinguishability in Genealogical Phylogenetic Networks at Lorentz Center* in Leiden (The Netherlands), in August 2016. The participants helped to organize the keywords into different sections to have different ways of exploring publications on the website: by network type, by goal type, by method type and by data type.

The website was regularly updated until 2020, when an unexpected change of version of PHP on the lab's server made the website unavailable. I made some necessary changes to upgrade it, including the BibAdmin free software it was based on (Che05), in August 2022. But so far I have not upgraded the code of the back-office part, and new research interests conducted me to stop updating the contents of the database.

1.1.2 Phylogenetic network classes

An Information system on inclusions of phylogenetic network classes

The keywords used in the website described in the previous section, illustrated in figure 1.3, can mainly be distributed among five categories: input data, software, studied problems, algorithmic properties and classes of phylogenetic networks. The latter corresponds to several mathematical restrictions which were introduced on the topology of phylogenetic networks in order to better describe some biological properties, to take into account limitations in reconstructability of the phylogenetic networks or to obtain faster algorithms. For example, unicyclic phylogenetic networks describe networks whose underlying undirected network contains exactly one cycle, which corresponds to modeling a genealogical history of species having experienced only one biological event of reticulation.

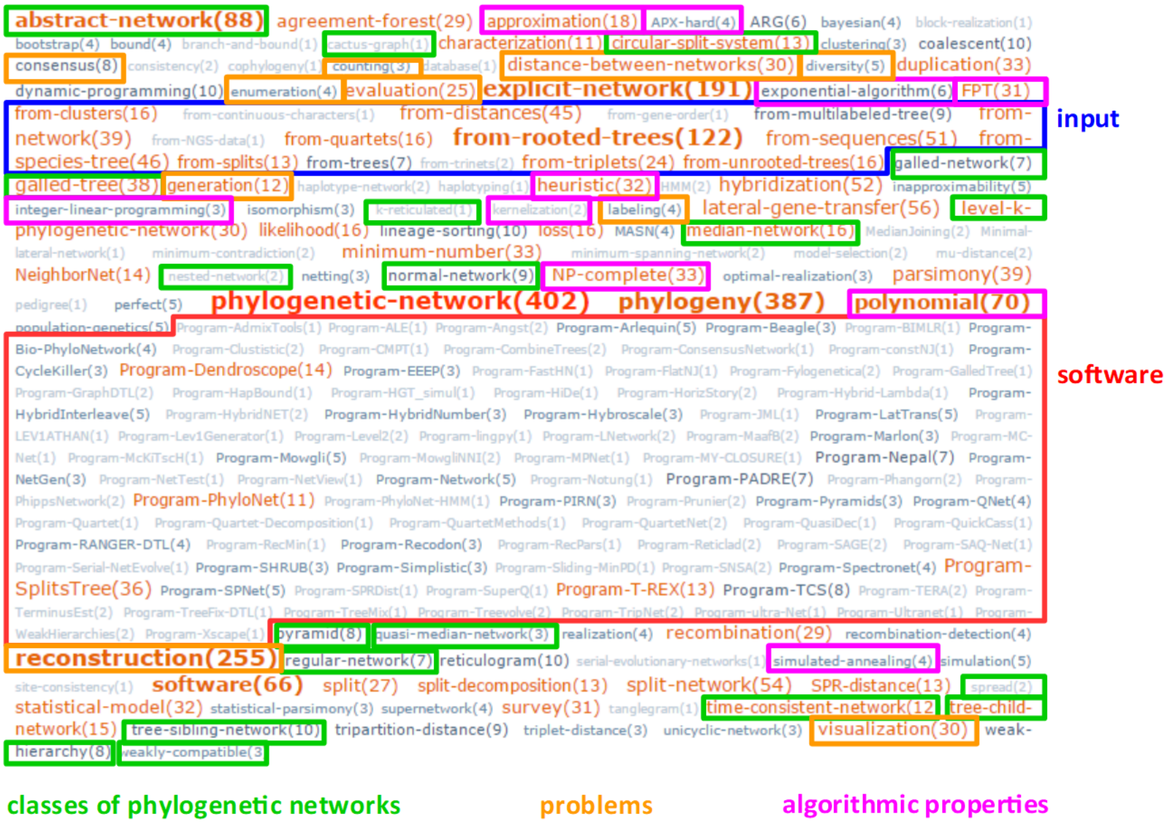


Figure 1.3: Word cloud of the keywords used to tag publications gathered in the website *Who is who in phylogenetic networks*, in 2015, colored by type of keyword.

Unfortunately, this system of tags labeling bibliographic references only provides limited information about the corresponding classes of phylogenetic networks. On the contrary, the website ISGCI, the *Information System of Graph Classes and their Inclusions* (dR$^+$), has been developed since 1999[1], after an idea by Andreas Brandstädt and Van Bang Le, as an encyclopedia on classes of graphs, to go beyond the book they had published with Jeremy Spinrad about graph classes (BLS99). Not only does it provide a mathematical definition of the restriction shared by graphs in the graph class, it also provides algorithmic results on the time complexity to solve several classical graph problems for this class. It also records inclusion relationships between classes of graphs, allowing to deduce algorithmic results from classes included in, or including, a given class.

Inspired by this online tool, I coded an equivalent for phylogenetic networks in PHP, *ISIPhyNC*, for *Information System on Inclusions of Phylogenetic Network Classes*[2]. Similarly, some problems are listed on this website, where NP-completeness can be deduced from NP-completeness on subclasses, and polynomial-time solvability can be deduced from the superclasses. Upper bound properties, such as on the number of nodes depending on the number of leaves, can also be deduced from superclasses. Furthermore, to prove non-inclusion

[1] https://www.graphclasses.org/

[2] http://phylnet.univ-mlv.fr/isiphync/

between classes of phylogenetic networks, examples of phylogenetic network contained in one class but not in another class are also provided.

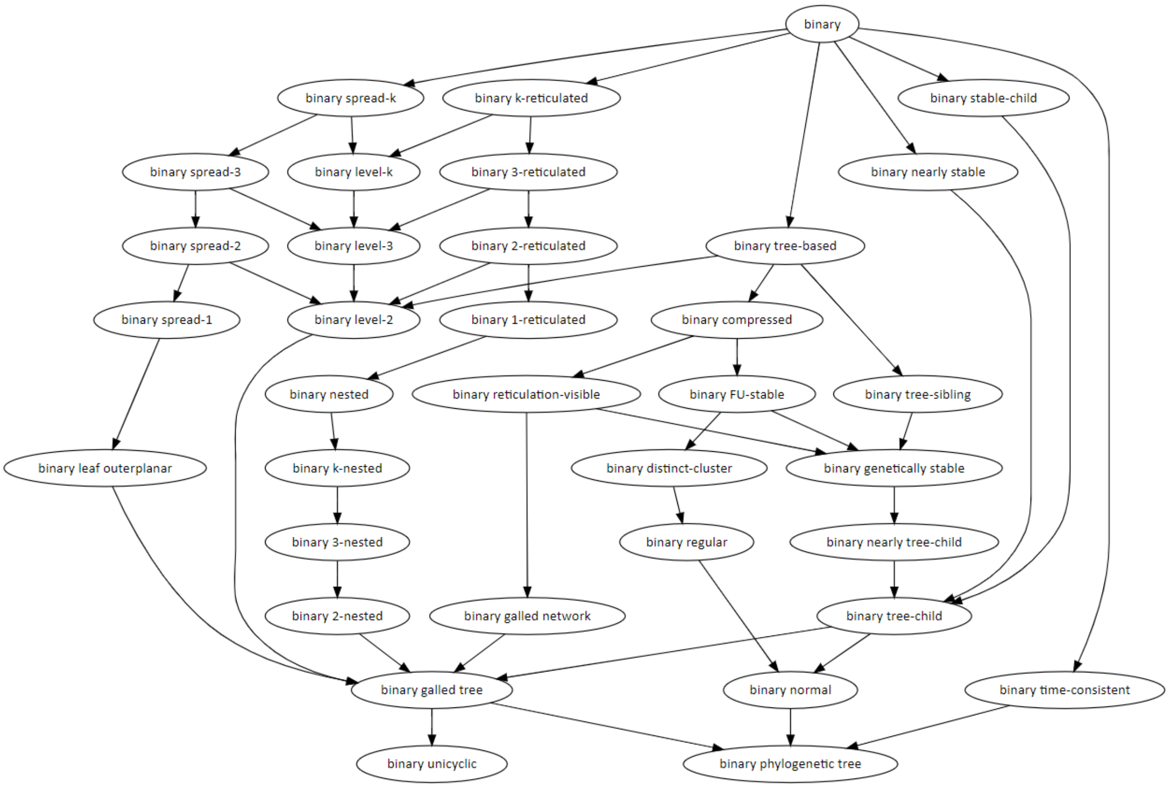
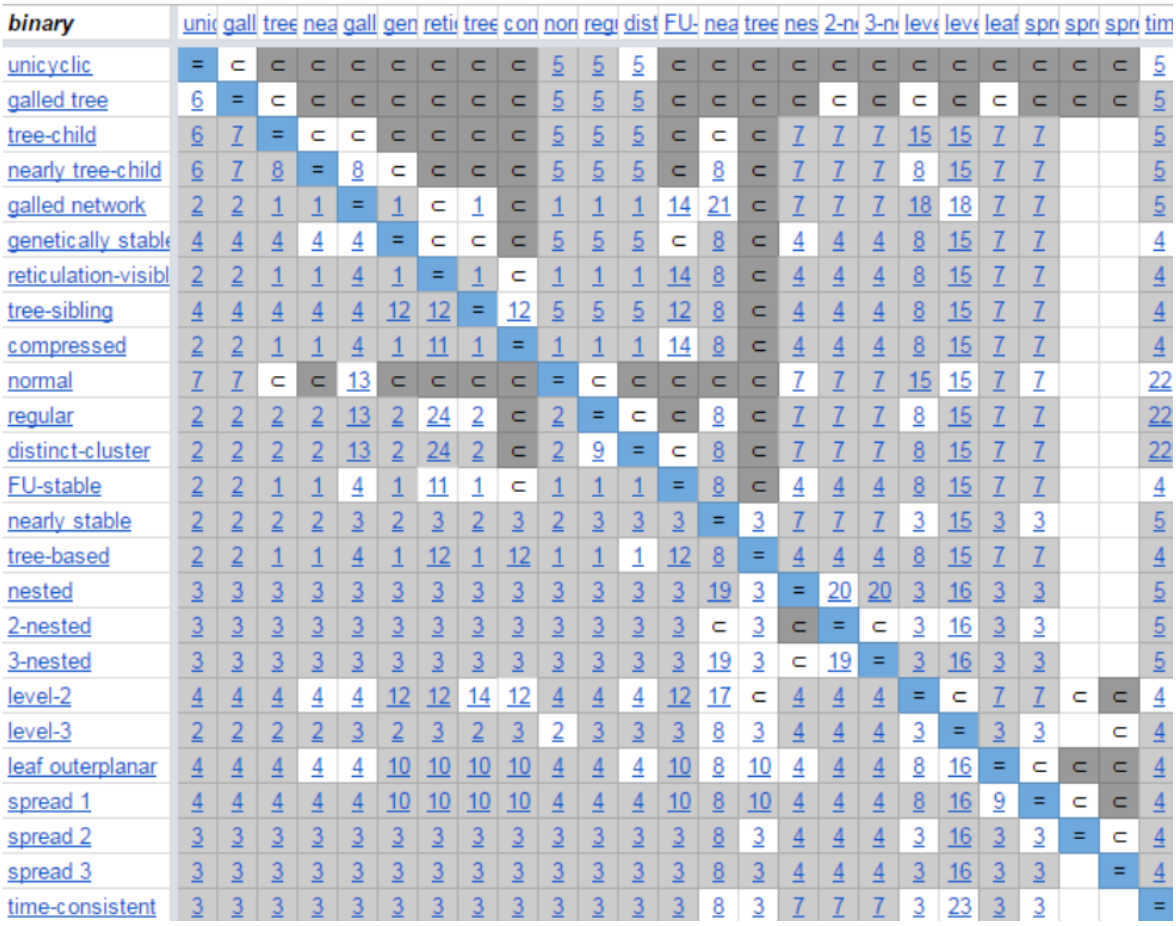


Figure 1.4: Inclusion relationships between classes of binary phylogenetic networks, from the *ISIPhyNC* website.

I coded the website during the Maxime Morgado's internship in June and July 2015, for his first year of master studies at ENS Cachan, when he started an exhaustive study on inclusions of classes of phylogenetic networks. To easily keep track of all inclusions and all open problems, I created a shared spreadsheet document[3], shown in figure 1.5, where all known inclusions between classes are represented, and if for two classes A and B, $A - B$ is known to be non-empty, an example of phylogenetic network in $A - B$ is provided.

This work was continued by Narges Tavassoli during a project for her bioinformatics course with me, for her second year of master studies at Université Paris-Est Marne-la-Vallée, from November 2016 to January 2017. The website currently contains 73 classes of phylogenetic networks including 35 classes of binary phylogenetic networks (defined in a total of 20 bibliographic references). 51 inclusion relationships were proved directly between those classes of binary phylogenetic networks (including some found in a total of 9 bibliographic references), as illustrated in figure 1.4. 24 phylogenetic networks can be viewed on the website, with a total of 68 proved membership to some class, and 56 non-membership to some class. For the three problems considered (Tree Containment, Cluster Containment and Phylogenetic Network Isomorphism) and for the three properties consid-

[3] https://docs.google.com/spreadsheets/d/1wtktAKdZ74u2MeDmeVLGHQnT3O0sfdy2NHVH4PAlchA/edit?usp=sharing



binary	unic	gall	tree	nea	gall	gen	reti	tree	con	nor	regu	dist	FU-	nea	tree	nes	2-n	3-n	leve	leve	leaf	spr	spr	spr	tim
unicyclic	=	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	5	5	5	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	5
galled tree	6	=	⊂	⊂	⊂	⊂	⊂	⊂	⊂	5	5	5	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	⊂	5
tree-child	6	7	=	⊂	⊂	⊂	⊂	⊂	⊂	5	5	5	⊂	⊂	⊂	7	7	7	15	15	7	7			5
nearly tree-child	6	7	8	=	8	⊂	⊂	⊂	⊂	5	5	5	⊂	8	⊂	7	7	7	8	15	7	7			5
galled network	2	2	1	1	=	1	⊂	1	⊂	1	1	1	14	21	⊂	7	7	7	18	18	7	7			5
genetically stabl	4	4	4	4	4	=	⊂	⊂	⊂	5	5	5	⊂	8	⊂	4	4	4	8	15	7	7			4
reticulation-visibl	2	2	1	1	4	1	=	1	⊂	1	1	1	14	8	⊂	4	4	4	8	15	7	7			4
tree-sibling	4	4	4	4	4	12	12	=	12	5	5	5	12	8	⊂	4	4	4	8	15	7	7			4
compressed	2	2	1	1	4	1	11	1	=	1	1	1	14	8	⊂	4	4	4	8	15	7	7			4
normal	7	7	⊂	⊂	13	⊂	⊂	⊂	⊂	=	⊂	⊂	⊂	⊂	⊂	7	7	7	15	15	7	7			22
regular	2	2	2	2	13	2	24	2	⊂	2	=	⊂	⊂	8	⊂	7	7	7	8	15	7	7			22
distinct-cluster	2	2	2	2	13	2	24	2	⊂	2	9	=	⊂	8	⊂	7	7	7	8	15	7	7			22
FU-stable	2	2	1	1	4	1	11	1	⊂	1	1	1	=	8	⊂	4	4	4	8	15	7	7			4
nearly stable	2	2	2	2	3	2	3	2	3	2	3	3	3	=	3	7	7	7	3	15	3	3			5
tree-based	2	2	1	1	4	1	12	1	12	1	1	1	12	8	=	4	4	4	8	15	7	7			4
nested	3	3	3	3	3	3	3	3	3	3	3	3	3	19	3	=	20	20	3	16	3	3			5
2-nested	3	3	3	3	3	3	3	3	3	3	3	3	3	⊂	3	⊂	=	⊂	3	16	3	3			5
3-nested	3	3	3	3	3	3	3	3	3	3	3	3	3	19	3	⊂	19	=	3	16	3	3			5
level-2	4	4	4	4	4	12	12	14	12	4	4	4	12	17	⊂	4	4	4	=	⊂	7	7	⊂	⊂	4
level-3	2	2	2	2	3	2	3	2	3	2	3	3	3	8	3	4	4	4	3	=	3	3		⊂	4
leaf outerplanar	4	4	4	4	4	10	10	10	10	4	4	4	10	8	10	4	4	4	8	16	=	⊂	⊂	⊂	4
spread 1	4	4	4	4	4	10	10	10	10	4	4	4	10	8	10	4	4	4	8	16	9	=	⊂	⊂	4
spread 2	3	3	3	3	3	3	3	3	3	3	3	3	3	8	3	4	4	4	3	16	3	3	=	⊂	4
spread 3	3	3	3	3	3	3	3	3	3	3	3	3	3	8	3	4	4	4	3	16	3	3		=	4
time-consistent	3	3	3	3	3	3	3	3	3	3	3	3	3	8	3	7	7	7	3	23	3	3			=

Figure 1.5: Overview of the spreadsheet document gathering all information about inclusion of classes of phylogenetic networks.

ered (upper bound on the number of nodes, unbounded number of nodes, completeness for reconstruction from trees), a direct proof is provided for 37 theorems (including some found in a total of 17 bibliographic references): 26 positive results can be extended to subclasses and 11 negative results can be extended to superclasses.

In 2022, Sungsik Kong and coauthors published a survey of classes of phylogenetic network, which focuses on 20 classes of phylogenetic networks, analyzing the biological or mathematical motivations to introduce them, as well as the big challenges on phylogenetic networks for which results were proven for some of these classes (KPKW22). In section 3 dedicated to inclusions between classes, they cite the ISIPhyNC website and mention it gives an “excellent overview” of inclusions between phylogenetic network classes. Their study contains 4 classes of phylogenetic networks which are not present in ISIPhyNC (orchard, valid, LGT and species-graph), and 7 inclusion relationships about these classes.

A collection of published phylogenetic networks

In order to be able to use an approach based on classes of phylogenetic networks in practice, I started to write a script in Python to test whether a phylogenetic network belongs to some of the most well-known classes of network. But best practices of software development including unit tests, code reviews and integration into a broadly used library to handle phylogenetic networks would be necessary to make such a tool really useful.

After discussions during the first Merlion project with Louxin Zhang, the idea of gathering a dataset of several networks published in research articles, to have a test set of realistic phylogenetic networks for various algorithmic applications, emerged. I therefore coded a webpage containing several thumbnails of pictures of phylogenetic networks extracted from scientific publications, shown in figure 1.6, linked with the bibliographic reference and with an encoding of the network, written manually, as a list of arcs. The list of arcs is generally established by manually performing a depth-first search of the network, adding labels along the way to the unlabeled vertices. The use of a Javascript implementation of GraphViz, a graph visualization software, allows to visualize the network online in a standardized manner, and a PhP script converts it into the eNewick format (CRV08), for use in other sofware[4].

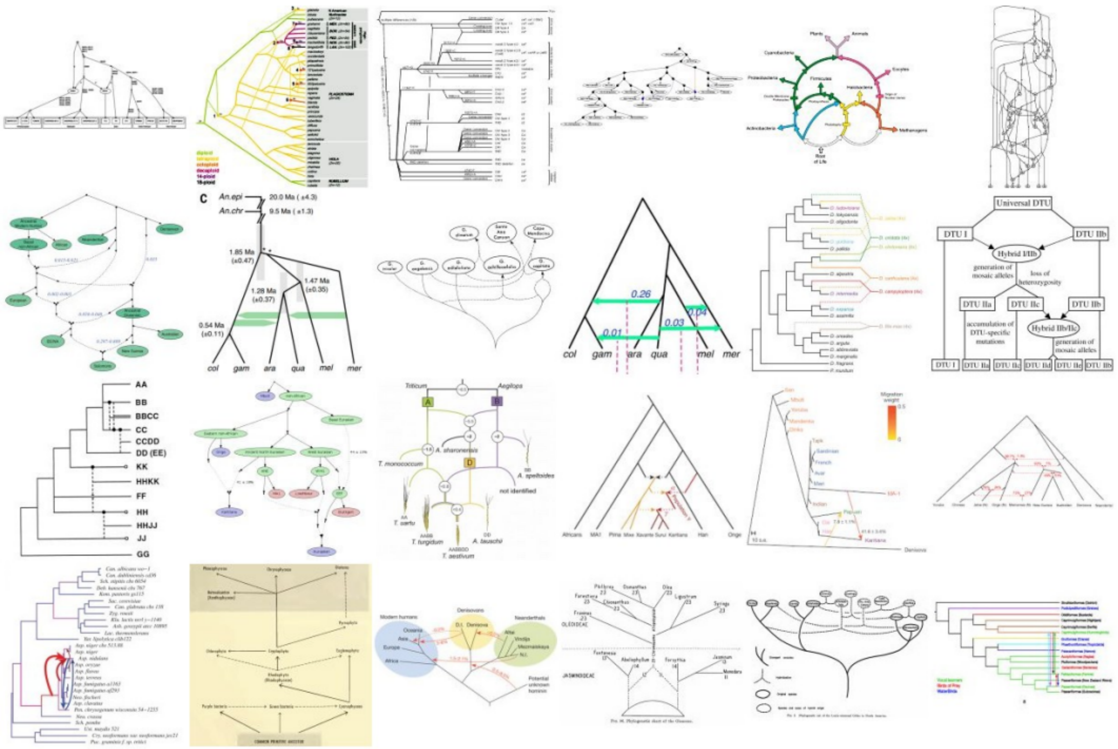


Figure 1.6: Overview of the collection of phylogenetic networks extracted from scientific publications. The bibliographic references of the sources of those networks are provided at https://phylnet.univ-mlv.fr/recophync/networkDraw.php.

This website provides a good activity for one-week interns in my lab, as adding networks and their encoding is a way of discovering not only the context of such networks in bioinformatics, but also the depth-first search algorithm, with a direct application and the ability to check with the resulting visualization if the resulting code is correct. It can

[4] The PHP code of this website is available at https://github.com/PhilippeGambette/networkDraw.

also be a way to write some HTML code to add the network to the webpage of the collection, and an opportunity to learn about best practice with research data, as this is the first dataset I have submitted to Recherche Data Gouv, the French national repository for research data (Gam23).

Therefore, adding encodings of networks for which I provided the picture was one of the tasks performed by high school students A. Sirunyan and S. Nguyen during their 3-day Science Académie internship at LIGM in 2016, as well as high school student Gabriel Schneider in December 2023.

1.1.3 Checking if a tree is contained in a network

On the way to Tree Containment

Several problems I studied during my PhD focused on the reconstruction of phylogenetic networks from a set of phylogenetic tree or some components of those trees (clusters, that is sets of leaves below some node of the tree, or triplets, that is topological restrictions of the tree to 3 leaves). The most basic problem for this approach is the Hybridization Problem which consists, given two rooted phylogenetic trees T_1 and T_2, in building a rooted phylogenetic network containing T_1 and T_2 and having the minimum number of hybrid nodes. More formally, we say the a rooted phylogenetic network N *contains* a rooted phylogenetic tree T having the same set of leaf labels as N if T can be obtained from N after a sequence of node deletions, arc deletions and arc contractions. The Tree Containment problem consists in deciding, when given a phylogenetic tree T and a phylogenetic network N having the same set of leaves, whether T is contained in N, as shown in figure 1.7.

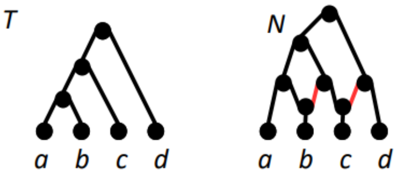


Figure 1.7: A rooted phylogenetic tree T contained in a phylogenetic network N (arcs are directed downwards), as T can be obtained from N by removing the two arcs in red and contracting the arcs just above leaves b and d and the two black arcs just above leaf c.

This model of tree containment is supposed to express the fact that even though hybridization, gene transfer or other biological events may explain that the genetic material of some species may come from two distinct species, it is likely that at gene level, each gene only comes from one parent species. Therefore, gene history is still tree-like, the gene trees being contained in the species network[5].

[5] The situation is actually more complex than expressed in this model: first, some genes, called mosaic genes, may contain components which come from two different parent species. Furthermore, due to gene duplication (and possiby deletion later), it is possible that two distinct co-existing copies of a gene are transfered together from a parent species: in this case, one arc of the phylogenetic network actually corresponds to two

As I had tried, during my PhD, to extend several results from the rooted to the unrooted context (GBP12), it made sense to try to extend the Hybridization Network problem to the unrooted context. It seems mathematically natural to define the problem in the unrooted context by simply replacing rooted trees and networks by unrooted ones, and by replacing the number of hybrid nodes of the rooted network by the *cyclomatic number* of the unrooted network, that is the minimum number of edges to delete to remove all circuits in the network. By proceeding in this way, we obtain the following Unrooted Hybridization Network problem: Given a set of input unrooted phylogenetic trees T_i and a positive integer k, does there exist an unrooted phylogenetic network N with cyclomatic number k which contains each of the input trees T_i ?

However, this question is not biologically relevant, as it is possible to build unrooted phylogenetic networks providing a positive answer to the Unrooted Hybridization Network, such that none of their rootings contain any rooting of the input trees. Therefore, they have little biological interest as they do not represent a possible evolutionary scenario.

A more biologically meaningful way of generalizing the Hybridization Network problem to the unrooted context is to define the Rootable unrooted hybridization network problem, in the following way. Given a set of input unrooted phylogenetic trees T_i and a positive integer k, does there exist an unrooted phylogenetic network N with cyclomatic number k for which there exists a rooting N' and a rooting T_i' for each input tree T_i such that N' contains each T_i'?

I proposed to study these problems for a "Merlion project" entitled *Two algorithmic issues of phylogenetic networks*, submitted in 2013 by Stéphane Vialette and Louxin Zhang, which was selected for funding by the French embassy in Singapore and the National University of Singapore, in 2014-2015.

Quickly after starting to work on these problems, also with Anthony Labarre and Andreas Gunawan, we realized that addressing the issues related with rooting in the variants of the Hybridization Network problem would require to get a deep understanding of the Tree Containment problem, which was proven to be NP-complete in (KNTX08), and for which a polynomial time algorithm was given in several restricted cases in (vISS10).

We note that the same issue about the generalization of the Hybridization network problem was identified by other authors, who published several results about it in 2018, namely on problems called Unrooted hybridization number and Root-uncertain hybridization number, which they proved to be NP-hard (VIKS$^+$18).

Tree Containment and reticulation visibility

We focused on special cases of Tree Containment for three classes of phylogenetic networks linked with the *visibility property* defined in (HRS11). A node n is said to be *stable* in a rooted phylogenetic network N if there exists a leaf $\ell(r)$ such that n belongs to all paths

arcs of the phylogenetic tree, so the tree is not contained in the network. In this case, when several arcs of the tree are allowed to be matched to the same arc of the network, we say that the network *weakly displays* the tree (HMSW16).

from the root to $\ell(r)$ in N. A phylogenetic network is defined to be *tree-child* if all nodes are stable (CRV09), *reticulation-visible* if all hybrid nodes are stable (HRS11), *nearly-stable* if each node is either stable or its parents are (GGL$^+$15a) and *genetically stable* if all hybrid nodes are stable and at least one of their parents is reticulation-visible (GGL$^+$15b).

The class of reticulation-visible networks was introduced in 2011 in (HRS11), where it is proved that solving the Cluster Containment problem, a problem similar to Tree Containment, can be done in polynomial time, for such networks. This class is a superclass of tree-child networks, for which the Tree Containment problem was proved polynomial-time solvable, it was therefore interesting to study if the problem would remain polynomial-time solvable for reticulation-visible networks.

Our first efforts focused on trying to bound the size of such networks, as explained in section 1.3. However, we were then not able to get a polynomial time algorithm directly on the class of reticulation-visible networks, only on nearly-stable binary phylogenetic networks, where similar bounds on the size of the networks were also obtained. For this class of networks, we show in (GGL$^+$15a) that focusing on the longest path from the root to a leaf ℓ allows to consider only a limited number of cases to decide, so far, if the input tree T still may be contained in the input network N, and if it is the case, reduce both T and N to continue the checking procedure on smaller instances. This results in a quadratic time algorithm to solve Tree Containment on nearly-stable binary phylogenetic networks. This result was improved by the authors of (FKP15) who provided an algorithm running in time $O(n \log n)$ thanks to a clever way of finding the longest path from the root to a leaf at each step.

We also provided a quadratic time algorithm to solve Tree Containment in binary genetically-stable phylogenetic networks (GGL$^+$15b) and obtained a linear time complexity algorithm for binary nearly-stable phylogenetic networks (GGL$^+$18).

The problem for reticulation-visible networks was solved in (GDZ17), using a new phylogenetic network decomposition technique and the algorithm was later improved to obtain a linear time complexity, independently in (Gun18) and in (Wel18).

Solving Tree Containment in practice: the database and the SAT approaches

We developed another approach to solve the problem in practice, after discussions started with Pierre Bourhis, a CNRS researcher in Lille. We initially wondered about connections between the trees which appeared in our domains, database theory for him, and more specifically XML-databases, where trees can represent the hierarchical structure of XML documents, and phylogenetic trees in my case. These discussions resulted in finding a paper about directed subgraph homeomorphism written in 1980 by Steven Fortune, John Hopcroft and James Wyllie, which provided a polynomial time algorithm to solve a problem closely related to Tree Containment (FHW80). We actually found out later that Juan Carles Pons also made this connection with directed subgraph homeomorphism in his doctoral thesis (PM16). Pierre Bourhis supervised a student at Université de Lille, Marion Tommasi, during a research project as well as a research internship in which I was also involved, to investigate these links further.

The Subgraph Homeomorphism problem consists in deciding, given two digraphs (directed graphs) P and G, and a one-to-one mapping m of the nodes of P into the nodes of G, if P is homeomorphic to a subgraph of G, that is deciding if there exists a mapping m' from the arcs of P to pairwise internal-node-disjoint paths in G such that each arc (t, h) is mapped to a directed path from $m(t)$ to $m(h)$, and such that all those paths are internal-vertex-disjoint (i. e. their only common vertices are their extremities).

Considering that the size of the phylogenetic tree in Tree Containment is constant, and that the size of the pattern P in Subgraph Homeomorphism is constant, Tree Containment reduces to Subgraph Homeomorphism in polynomial time.

Indeed, given an instance of Tree Containment, that is a binary phylogenetic tree T with a constant number k of leaves and a binary phylogenetic network N with k leaves (with the same labels as T's) and with n vertices, T is contained in N if and only if there exists a mapping m from the root of T to the root of N, from the leaves of T to the leaves of N with the same labels, and from the other vertices of T to vertices of N such that the Subgraph Homeomorphism problem has a positive answer for T, N and this mapping m. Therefore, this instance of Tree Containment has a positive answer if and only if one of the $O(m^k)$ mappings of vertices of T into vertices of N, combined with the relevant mappings of the roots and of the labeled leaves, has a positive answer for the Subgraph Homeomorphism problem. If k is a constant, this can be tested in polynomial time using the pebble algorithm described in section 4 of (FHW80).

This idea was developed by Marion Tommasi, into a Datalog program and later in an optimized Python script she implemented[6]. Unfortunately, the first approach where all possible matchings had to be tested with the pebble algorithm[7] was too slow. Therefore, new strategies were designed, inspired by classical optimizations of Datalog programs, to take advantage, in the pebble algorithm, of the matchings which are actually already known from the start: between the leaves having the same label in the tree and in the network (bottom-up approach), and between the root of the tree and the root of the network (top-down approach). With such strategies, matchings between the nodes of the tree and the nodes of the networks are gradually being built, and as we stop building them as soon as the pebble algorithm detects an invalid pebble configuration, we avoid considering some of them which will never extend into a valid matching between nodes of the tree and the network. Other optimizations introduced to fasten this approach include grouped pebble moves (when a pebble has to continue moving and has to move along one arc of the network) and parallelization.

Another approach consists in using a SAT solver, as we proved that the Tree Containment problem can be reduced to 5-SAT in polynomial time. I implemented an algorithm[8] which, given a phylogenetic tree T and a phylogenetic network N with n vertices as input, builds an instance $C(N, T)$ of the 5-SAT problem, in time $O(n^3)$.

[6] Available at https://gitlab.inria.fr/Spirals/logical-approach-for-tree-containment

[7] In this algorithm, pebbles, which are moved along the arcs of the network, represent matchings between an arc of the tree and an extremity of a path in the network which may be a candidate, so far, to be matched with this arc.

[8] The Python code, which also makes an external call to the call to the Sat4j solver (LBP10), is available at https://gitlab.inria.fr/Spirals/logical-approach-for-tree-containment

The general idea is to build this formula using variables x_{a_T,a_N}, where (a_T, a_N) is a pair of arcs of the tree and the network respectively, such that x_{a_T,a_N} has value *true* if $m(a_N) = a_T$, *false* otherwise. The fact that a_N is mapped to a_T has consequences on the possible mappings of arcs just below a_N, which leads to building clauses on the variables depending on local configurations of arcs below a_N and a_T. Finally, it can be proven that the disjunction of all clauses built in this way is true if and only if T is contained in N.

Some preliminary tests run by Marion Tommasi showed the complementarity of the "database approach" and the "SAT approach", for 100 instances of the Tree Containment problem, 10 positive ones and 90 negative ones, built in the following way:

- 10 random networks with 100 leaves and 100 hybrid nodes were generated with the random phylogenetic network generator of (Zha16);
- 10 trees on 100 leaves, each extracted from one of the networks in the following way were obtained by randomly deleting, for each hybrid node, an arc coming from one of its two parents.

On all the negative instances, the "database approach" was always faster, taking less than 0.01 second to ouput an answer. However, on positive instances, the "SAT approach", which required approximately 55 seconds for formula generation (files of approximately 150 Mb were generated for each instance), and 7 seconds for the Sat4j solver to provide an answer, was sometimes faster than the "database approach".

This encouraged us to push further this practical analysis of the Tree Containment problem, which is ongoing work with Sarah Berkemer. I presented part of this work in my invited talk at the workshop at the Institute for Mathematical Science of the National University of Singapore in September 2023 (BBG+23a).

Tree Containment and branch lengths

In 2016, I was invited by Celine Scornavacca, a CNRS researcher in Montpellier who defended her doctoral thesis in Montpellier, also under Vincent Berry's supervision, to be involved in her PICS project, funded by the CNRS, to work with Fabio Pardi, also based in Montpellier, and with colleagues in the Netherlands, Leo van Iersel, Steven Kelk, about practical approaches to reconstruct phylogenetic networks.

In the search of algorithms capable of handling more realistic models and more complex input data, we wondered if the Tree Containment problem would be easier to solve if the branch lengths of the input trees were provided, hoping that this extra information might help to know where to match the nodes of the tree inside the network.

In (GvIJ+17), we proved that the problem we called Tree Containment with Branch Lengths is strongly NP-complete for binary phylogenetic trees and networks, even if the networks are tree-sibling[9] and time-consistent[10], reducing from the 3-Partition problem.

[9] A rooted phylogenetic network is *tree-sibling* if every hybrid node has at least one sibling which is not a hybrid node (Nak04).

[10] A phylogenetic network N is *time-consistent* if each node v of N can be labeled with an integer $t(v)$ such

A reduction of the Subset Sum problem, where we want to decide whether it is possible to obtain a given input value by summing together some of the integers given as input, shows that the problem is weakly NP-complete for level-k binary phylogenetic networks[11], for $k > 2$. Given an instance of the Subset Sum problem, with input integers n_1 to n_k, we show how to create an instance of the problem Tree Containment with Branch Lengths where the network is built by chaining together directed acyclic graph gadgets with arc lengths such that the path of the tree possibly contained in the network has to go through each gadget and choose either a path having length 2, or a path having length $n_i + 2$. Therefore, reaching the desired number s as a sum for the Subset Sum problem is possible if and only if it is possible to reach a total length of $s + 2k$ for the part, intersecting these gadgets, of the path corresponding to an arc of the tree T contained in the network N.

Note that a key component in this reduction is that the path going through all those chained gadgets in the network is forced to correspond to only one single arc of the tree contained in the network, which is possible because each gadget is a bridgeless component with only one outgoing arc in the phylogenetic network. We call those *redundant* bridgeless components.

Then, we show that the problem Tree Containment with Branch Lengths is fixed-parameter solvable in k, for level-k networks with no redundant bridgeless components. For this, we use a bottom-up algorithm using brute-force to check consistency with the lengths of all possible paths inside the bridgeless components, simply extending the algorithm designed for level-k networks with unknown branch lengths. This strategy can also be generalized to provide an algorithm to solve the problem in pseudopolynomial time when the network contains redundant bridgeless components.

Playing Tree Containment with junior high school students

In 2017, I designed a device to communicate about this research work during the "Fête de la Science"[12] It consists in a board of wood, where a printed phylogenetic networks was glued. A nail is positioned on each internal node of the network, different kinds of nails being used for hybrid nodes and tree nodes. The players are then provided with an unrooted phylogenetic tree made of elastics, where the leaves are replaced by key rings labeled by the leaf label. The goal is then to find out whether the tree is contained in the network by respecting the following rules, which correspond to the Tree Containment problem: the keyrings must be attached to the nails corresponding to their label, the elastic branches of the tree must follow the branches of the network (the nails help to keep them in place)

that a hybrid node has the same label as its parent and any other node has a label strictly greater than its parent (BSS06).

[11] A phylogenetic network, rooted or unrooted, has *level-k* if it is possible to obtain a tree by removing at most k arcs or edges from each bridgeless component (a *bridgeless component* of a network N is a subset C of vertices of N such that the subnetwork of N induced by C is connected and does not contain any bridge, that is an edge whose removal disconnects the graph).

[12] National science week occurs every year in October in France. Animations for junior high school students of the area are usually organised on Wednesday by Ifis, one of the teaching departments of Université Gustave Eiffel.

and two elastic branches cannot be positioned over the same branch of the network and starting from the root of the network, the elastic branches must always go downwards.

When we first organized this game in 2016, Anthony Labarre and I gave some details about the meaning of the tree and the network, using a poster prepared for the event. After the students had found the solution, they would win a Carambar[13]. On the next two years, Laurent Bulteau had another game prepared next to mine, where he would make students discover the BUILD algorithm (ASSU81) to reconstruct a phylogenetic tree from a subset of its triplets. This was a good oportunity to make lasting impressions on students who experimented with both games.

1.1.4 Exploring the space of rooted phylogenetic networks

As explained above, the research conducted during the PICS project supervised by Celine Scornavacca aimed at developing more practical methods for phylogenetic tree reconstruction, which encouraged us to focus on recent results about local rearrangement operations for phylogenetic networks.

Indeed, for several reconstruction methods aiming at finding an optimal phylogenetic tree from biological sequences associated with their leaves, as this optimization procedure is NP-hard for several optimization criteria, such as maximum parsimony (NJZMC05) or likelihood (JNST06), a local search heuristic is often a fruitful strategy in practice. It works by gradually improving the phylogenetic trees reaching the best scores, by modifying them slightly. To this aim, several kinds of local moves are used, for example nearest neigbhor interchange moves (NNI moves for short), which consist in exchanging the subtrees attached to the two extremities of one edge (or arc, in the rooted case) of the tree.

Local moves called LST operations had been defined on unrooted and rooted phylogenetic networks in (HLMW16): they included an equivalent of NNI moves to transform a phylogenetic network into another one with the same level, three-cycle pop operations to obtain a network of upper level and three-cycle shrink operations to obtain a network of lower level. The authors proved that it was possible to obtain any unrooted (respectively rooted) level-1 network from any other unrooted (resp. rooted) level-1 network using LST operations. of reticulate evolutionary histories.) In (HMW16), using similar operations (NNI moves, δ^+ and δ^- operations which extend three-cycle pop and shrink to any un-

[13] Jokes are usually written on Carambar wrappers. Carambar was a natural choice as I had written two blog posts about methods to evaluate the total number of distinct jokes on Carambar sweets, or the number of distinct quotes around “papillotes”, chocolates eaten at Christmas in France, during my PhD (Gam09). Several statistical methods, with various degrees of difficulty, can be described, an easy one inspired by the capture/recapture method to evaluate the size of a population of wild animals, a more difficult one based on estimating the number of repeated jokes after eating n Carambars. This mathematical story about Carambars and papillotes was featured in an article in the French newspaper *Le Monde* on March 27th, 2013 after the journalist Sandrine Blanchard found my blog post online (Bla27), and adapted in a mathematical popularization book written by Jérôme Cottanceau (Cot16). I also adapted it into a “street mathematics” show during the “Science en marche” protests in 2014, which demanded better fundings for research in France: I presented those methods during a crowdsourced experiment on Carambar joke counting, in a stand on the streets of Champigny-sur-Marne, next to other science activities proposed by colleagues of my university.

rooted network) it is proven that any unrooted phylogenetic network can be obtained from an unrooted phylogenetic tree by a sequence of those local moves, and that it is possible to obtain any unrooted phylogenetic network from a phylogenetic network with the same number of nodes by a sequence of NNI moves. In the discussion section, the authors suggest that based on previous work, extending these results to rooted phylogenetic networks "will probably be somewhat more technical" but "it seems to be well worth while trying to understand general properties of rooted network spaces".

This motivated the study of a new definition of *NNI* moves for rooted phylogenetic networks, which we introduced in (GvIJ+17). It extends the definition of NNI moves on rooted trees and is consistent with the definition of NNI moves on unrooted networks introduced in (HLMW16). We also introduced another rearrangement move called rooted *SPR* (subnetwork pruning and regrafting), extending the SPR (subtree pruning and regrafting) moves initially defined for phylogenetic trees (Fel03). We proved that for rooted networks, NNI moves are special cases of SPR moves, and that it is possible to search the whole space of binary phylogenetic networks with the same set of leaves also having the same number of nodes (or equivalently, the same number of arcs, or the same number of hybrid nodes) by a sequence of rooted NNI moves.

My coauthors made further progress on the subject with other colleagues. For example, on the theoretical side, they noticed that among rooted NNI or SPR moves, only some of them, called *tail moves*, were really useful to ensure connectivity among binary phylogenetic networks having the same set of leaves and the same number of hybrid nodes, they proved that the resulting distances between phylogenetic distances are NP-hard to compute and also estimated lower and upper bounds for the diameter of those spaces of phylogenetic networks. On the practical side, a computer program called NetRAX was developed to implement a fast and efficient maximum likelihood method to reconstruct phylogenetic networks using these rearrangement moves for the local search (LSK+22).

1.2 Characterizing and identifying a phylogenetic network with its subparts

1.2.1 Using complete information

Several approaches were developed to reconstruct phylogenetic networks from some finite sets of elements extracted from them, namely clusters, splits, triplets, quartets and trees. To help designing efficient reconstruction methods, one can wonder if these sets of elements may have special properties for some classes of phylogenetic networks. Another interesting question is whether phylogenetic networks can be identified from such sets of subparts: if several phylogenetic networks have exactly the same set of contained trees, outputing only one solution after a network reconstruction algorithm taking trees as input may provide the wrong network.

My doctoral thesis included a few results about these two issues. Some of them were obtained with my supervisors Vincent Berry and Christophe Paul, published in 2012 (GBP12):

for unrooted level-1 networks, we proved that their *split system* (that is, the set of all bipartitions of leaves induced by minimal cuts in the network) is *circular*, meaning that it is possible to order the leaves of the network so that one of the sets of this bipartition is an interval (that is, its leaves are consecutive) in this order. With Katharina Huber, from University of East Anglia, we had focused on the uniqueness issue (GH12). We described a subset of rooted level-1 networks, namely, the ones having no bridgeless component of 3 or 4 nodes, where it is possible to identify networks from their sets of contained trees, triplets or *clusters* (sets of leaves below each node). We also showed that for level-1 networks outside this subset, it is always possible to find a distinct level-1 network having exactly the same set of contained trees, triplets or clusters.

We continued working on these topics with Katharina Huber after my doctorate, in particular when she visited me in Marseille in 2011, thanks to a grant from the London Mathematical Society. As she was one of the authors of a book about abstract phylogenetic networks and splits (DHK$^+$11), her extensive knowledge of this topic gives her another perspective on such mathematical objects and made scientific discussions on these problems very exciting. In particular, I remember a walk to the calanques where we had the idea of the intersection closure to characterize the set of splits associated with an unrooted level-1 network, which eventually resulted in the first part of the theorem 1 of (GHS17). Our first results were later completed with the contribution of Guillaume Scholz, who added a few results to this paper as part of his doctoral work with Katharina Huber.

1.2.2 Using partial information

In (GHK17), we continue to study whether level-1 phylogenetic networks can be unambiguously reconstructed from triplets and clusters, but we obtain results on cases where this is possible from a strict subset of well-chosen triplets. In particular, we present a subclass of level-1 networks where each member can be reconstructed unambiguously, in this subclass, using a subset of at most $2n - 1$ its triplets, where n is the number of leaves of the network.

1.3 Counting phylogenetic networks and their components

1.3.1 Counting level-k networks

When I presented the paper about the decomposition of rooted or unrooted level-k networks at CPM 2009 in Lille, where Mathilde Bouvel, an expert in combinatorics in general and enumerative combinatorics in particular, was also present, we thought that the tree-like structure of level-k networks would probably make it possible to count them. In 2006, Semple and Steel had already published a paper about counting unrooted level-1 network (SS06). Mathilde Bouvel obtained a grant from her lab, Labri in Bordeaux, to invite me for a week, during my postdoc, in 2011, to work on it together.

Using the analytic combinatorics framework developed by Philippe Flajolet and Robert Sedgewick (SF09), we used the tree-like structure of rooted and unrooted level-1 and unrooted level-2 networks to write recursive descriptions of these objects, the technical part being to pay attention to symmetries, therefore avoiding to describe automorphic networks several times. These descriptions allowed us to obtain exact and asymptotic enumeration formulas to count those networks given their number of leaves, and possibly also their number of cycles and edges across the cycles.

I presented those results in a talk at the Newton Institute and at the East Anglia University in Norwich. However, writing the paper about these results was quite problematic: the initial idea was to write an extended journal version of the CPM 2009 paper mentioned above. But to be able to consider that the components of level-k networks, called level-k generators, were also level-k networks themselves, the definition of phylogenetic networks we introduced in this paper allowed for example multiple arcs, as well as hybrid nodes to have outdegree 0, therefore also being leaves. Combining the previously obtained results about such "generalized" level-k networks, and the links between the rooted and unrooted version, and the new counting results on traditionally-defined level-k networks, required to clarify everything more formally. However, a 2016 paper introduced the analogue of level-k generators for unrooted networks (HLMW16), which made this contribution about the structure of rooted and unrooted level-k networks no longer relevant, and encouraged us to focus on counting problems.

In May 2018, Marefatollah Mansouri, who was preparing his doctoral thesis at TU Wien, where he was also working on phylogenetic network counting problems with his advisor Bernhard Gittenberger, contacted me about these results. We decided to join forces to finish this paper, focusing only on counting results, adding results on rooted level-2 networks and following the advice of the reviewers of the first submitted version of the paper, to focus on the counting of traditionally-defined rooted and unrooted level-1 and level-2 networks. A summary of the obtained results is given in table 1.1. Bernhard Gittenberger also invited me to give a talk in Vienna, during the meeting of the ANR-FWF-MOST project, and I was happy to see, in the *Workshop 1: Foundations of Networks*, organized by Daniel Huson and Louxin Zhang at the Institute for Mathematical Science in Singapore, in September 2023, how much progress had been made on the combinatorial description and enumeration of various classes of phylogenetic networks.

1.3.2 Bounding quantities related with phylogenetic networks

Diameters for distances between networks

Counting combinatorial objects has also been involved in other results about phylogenetic networks. With Katharina Huber, when we studied the possibility of uniquely encoding phylogenetic networks by trees, clusters or triplets (GH12), questions about the corresponding distances naturally emerged. Indeed, one way to measure how similar two phylogenetic trees, or networks having the same set of leaves, are, is to compute the size of the symmetric difference of their family of trees, clusters or triplets. The smallest it is, the most similar we expect them to be, as their topologies share a lot of common components. In this context,

n	g_{n-1}	r_n	u_{n-1}	ℓ_n
1	0	1	0	1
2	1	3	1	18
3	2	36	6	1 143
4	15	723	135	120 078
5	192	20 280	5 052	17 643 570
6	3 450	730 755	264 270	3 332 111 850
as $n \to \infty$	$c_1 \approx 0.20748$	$c_1 \approx 0.1339$	$c_1 \approx 0.07695$	$c_1 \approx 0.02931$
$x_n \sim c_1 c_2^n n^{n-1}$ with	$c_2 \approx 1.89004$	$c_2 \approx 2.943$	$c_2 \approx 5.4925$	$c_2 \approx 15.4333$
OEIS reference	A328121	A328122	A333005	A333006

Table 1.1: The numbers g_{n-1} and u_{n-1} of unrooted networks, of level-1 and level-2 respectively, on n leaves and the numbers r_n and ℓ_n of rooted networks, of level-1 and level-2 respectively, on n leaves, as well as the identifier of the sequence describing them in the *Online Encyclopedia of Integer Sequences* (Slo94).

the characterization question presented in section 1.2 can be rephrased as: given a phylogenetic network N, does there exist another network N' such that the distance between N and N' is 0?

We therefore studied the diameter of these distances for the class of phylogenetic networks we had focused on, namely level-1 networks whose underlying undirected graphs do not have cycles of length smaller than 4. We proved that for networks of size n in this class, the diameter of the tree distance is $2^{\lfloor \frac{n-1}{3} \rfloor}$ and the diameter of the cluster distance is $2n - 4$.

Bounded sizes of networks

When we focused on special classes of phylogenetic networks to study the Tree Containment problem presented in section 1.1.3, we also had some counting to do, on the maximal size of a network in the phylogenetic network classes we focused on, depending on its number of leaves.

It happens that the size of those three classes can be bounded when their number n of leaves is bounded. In (GGL$^+$15a), we proved upper bounds for the number of hybrid nodes of reticulation-visible networks ($4n - 4$) and nearly-stable networks ($12n - 12$).

These results were later improved in (BS16) to obtain a tight upper bound of $3n - 3$ for reticulation-visible networks in (GZ15) to obtain a tight upper bound of $3n - 3$ for nearly-stable networks.

Chapter 2

Exploring Data Based on Proximity

2.1 Studying two proximity parameters in graphs, contiguity and modularity

2.1.1 Graph contiguity: forcing proximity in the neighborhoods

My doctoral dissertation included a summary of some results I had obtained with Christophe Crespelle about a new graph parameter, graph *linearity*, which was proposed by Christophe to take advantage of the fact that neighborhoods in graph may be represented as unions of intervals in a set of orders of the vertices (one interval chosen in each order). This parameter was introduced in (CG09) as a variant of the *contiguity parameter*, the lowest k such that the adjacency matrix of the graph has the k-consecutive ones property, introduced in (GGKS95), meaning that it is possible to find an order on the vertices of the graph such that the neihbourhood of each vertex is equal to a union at most k intervals in this order.

Our article had however left some questions unanswered, and Christophe proposed to co-supervise the bachelor internship of a student from ENS Lyon, George Manoussakis, to extend our work on this topic. George worked on those two parameters on trees, where he gave a linear-time algorithm to compute them. He also worked on cographs, finding the value of contiguity for cographs having a complete binary tree. We therefore tried to obtain further results on cographs. In particular, we used tree decomposition techniques, related with the *rank* parameter on trees, to show how to obtain upper and lower bounds for the value of contiguity on cographs: the contiguity is at most 2 rank(T) +1, where T is the cotree associated with the cograph, and at least (rank(T)-7)/4, which actually provides a 23-approximation algorithm running in linear time, to compute the contiguity of cographs (CG13).

We also worked on similar results for the linearity parameter. But contrary to contiguity, for which the $O(n)$ upper bound is asymptotically tight (as the contiguity of cographs with complete cotree is $\Omega(\log n)$), we only obtained an asymptotic lower bound of $\Omega((\log n)/(\log \log n))$, and the same $O(n)$ upper bound, for linearity (CG14). Christophe Crespelle was later able to improve this result with other coauthors, proving that the linearity of cographs has an asymptotic upper bound of $O((\log n)/(\log \log n))$, therefore showing that linearity is strictly more powerful than contiguity for encoding graphs (CLPP16).

2.1.2 Balancing speed and precision for modularity-based graph clustering

In 2010/2011, I spent nine months as a postdoctoral researcher with Alain Guénoche for the ANR Moonlight project supervised by Christine Brun, which aimed at identifying moonlighting proteins, that is proteins having several functions in the cell. The chosen approach was to design a variant of a graph partitioning algorithm allowing to output overlapping clusters, and to run it on protein interaction networks: the vertices belonging to 2 overlapping clusters may be good candidates to correspond to moonlighting proteins, having one function for each computed protein cluster containing it.

A fast and efficient method to detect communities in networks is the Louvain method, a heuristic which aims at optimizing the *modularity* of the output partition (BGLL08). The modularity score aims at estimating the surplus of density inside the subgraphs induced by the partition, compared with a situation where edges would be distributed uniformly in the network. As protein interaction networks typically have thousands of vertices, a new, slower but more robust method was designed, obtaining better modularity scores. We therefore kept the agglomerative hiearchical clustering principle of the Louvain method, where clusters are transfered from one class to another one as long as modularity increases, before merging the clusters of the same class into the clusters of the next step. But for each of these steps, transfers of individual vertices of the input network are also tried between classes, keeping the transfers which improve modularity. This algorithm, called TFit, for "transfer and fusion…iterated", can be adapted to allow overlapping between classes. Its robustness to noise in the data can also be improved using a bootstrapping procedure (GG11). This procedure creates several slightly altered copies of the input network, where some randomly chosen edges were modified, to run the TFit algorithm on them, and then combine all obtained partitions through a consensus method designed by Alain, actually a heuristic similar to TFit (Gué11).

I coded the TFit algorithm in Java starting from Alain's code in C, in order to be able to include it as a plugin in the Cytoscape software, highly used for various kinds of network analysis in bioinformatics. With the help of Laurent Tichit, a proof of concept of Cytoscape plugin was prepared and presented at MARAMI 2011 in Grenoble (TGG11). I also compared it with other graph partitioning algorithms on a benchmark of partitioned networks. But as my postdoc was interrupted before its initially planned conclusion when I was hired by Université Paris-Est Marne-la-Vallée in September 2010, the integration into Cytoscape with a nice easy-to-use plugin was coded by Lionel Spinelli after I left Marseille (SGC$^+$13).

This work was also extended by a collaboration with Marie-Hélène Mucchielli-Giorgi and Annie Glatigny, for the development of an automatic identification method of protein sub-complexes in a multiprotein complex (GGBP$^+$17). We designed an ad-hoc method which obtained more robust results than the modularity-based method TFit and other network partitioning methods, for this goal. More precisely, an agglomerative clustering method was designed, based on the similarity of proteins contained in the complex, computed as the similarity of their neighborhoods in the protein interaction graph.

Several formulas exist to express the similarity between two sets of elements, which I had explored in my work on tree clouds (see section 2.2). The simplest one, the Jaccard formula, was actually the one which happened to be the most consistent with the biological results. Note that for two proteins a and b, this score is not computed exactly on the neighborhoods of a and b in the network, but on their neighborhood *outside of the complex* in the network, that is ignoring all edges between protein in the complex. This adaptation illustrates the importance of taking into account the biological context (in this case, deciding that the important interactions for sub-complex identification are actually the ones with proteins outside the complex) when designing a method inspired by another field (formulas of similarity used to express proximity between words in a text).

For this project, an R package, called *ISIPS*[1], was designed and coded by Marie-Hélène Mucchielli-Giorgi, with a few contributions from my side.

2.2 Tree clouds: statistical textual analysis based on word proximity

I discovered a tree of words for the first time on Jean Véronis's blog. This tree was built to visualize the proximity between those words in the text, using a phylogenetic tree reconstruction algorithm. This encouraged me to program TreeCloud, a software to built such representations, and to contact him and suggest we write an article together about the choice of parameters of the chained algorithmic components, to obtain trees which would be robust to small changes in the text (GV09). My first contribution outside the field of bioinformatics and graph algorithms, was the starting point of many exciting collaborations with researchers in other fields, both during my doctorate (AG10; LGMT10) and after; it has since then become my most cited article.

2.2.1 Using and advertising TreeCloud

Our 2010 communication with Delphine Amstutz at the JADT conference, which illustrated how to use TreeCloud in combination with other tools of statistical textual analysis, for the comparative analysis of two plays by Corneille, actually paved the way for the methodology I developed after my doctorate, on how to use tree clouds for textual analysis. We had indeed noticed that it could be used either to generate exploratory hypotheses about the corpus (typically, by identifying the main topics in the corpus, or by identifying similarities and differences between the tree clouds of two subcorpora), to confirm hypotheses (typically, by building the tree cloud of a subcorpus of interest, such as the speeches of one character of interest in a play, or of the contexts of one word of interest), or to represent the results of other statistical tools for textual analysis (for example, to display specificity score with word sizes in the tree cloud, instead of inside a table of specifity scores).

Some discussions with Edna Fernandez, a postdoc working on analyzing perceptions of the urban environment during night walks in cities, helped to design a more precise list of use cases of TreeCloud which I used in the presentation of TreeCloud I have given annually in the doctoral seminar of Jean-Marc Leblanc at Université Paris-Est Créteil, from 2014 to 2023, and in the master courses about text mining by Renaud Epstein at Université Gustave Eiffel, since 2016. Each use case is illustrated by an example coming from one of the publications summarized below, and I sometimes replace one figure by a more impactful one, for the sake of clarity.

In (GM13), with William Martinez, whom I had met in a workshop in Besançon, where I presented the work detailed in the next section, we analyzed a corpus of newspaper articles about the Mediator drug which caused a health scandal in France after being prescribed

[1] http://bim.i2bc.paris-saclay.fr/isips/

outside of its original scope, causing several deaths or severe illnesses. One goal of this article was to compare articles written by news agencies in particular, with articles written by journalists who did not mention news agency as their main source. When comparing the tree cloud of articles by news agencies with the other ones, it was noticeable that the semantic clusters seemed more focused in the latter, whereas the main topics clearly identified in the general tree cloud for the whole corpus were scattered in different subtrees in the corpus of articles by news agencies. Our hypothesis was that contrary to targeted articles by journalists who would focus on one topic, and therefore increase the cooccurrence relationships between words about this topic, news agencies might give a broader description of the story, including context which would add cooccurrence relationships between words linked with different topics. The tree cloud of the contexts of the word "responsabilité" in both subcorpora also suggested that the articles not written by news agencies had paid more attention to the role of doctors prescribing Mediator.

A project with members of LISIS (another lab of my university) to analyze how scientists construct an argument in the abstracts of research projects about biodiversity (CBB+14) led to a collaboration with Xavier Le Roux, who supervised the BiodivERsa project. The goal was to build and analyze the tree clouds of the abstracts of research projects about biodiversity, funded by research agencies in Europe and listed in the BiodivERsa database, to uncover temporal and geographical trends. It was especially interesting to see Xavier Le Roux apply the methodology to build relevant tree clouds, first by removing words which could be considered stop words in this context, such as "research" or "projects". Then he obtained visualizations which confirmed some intuitions he had about differences between projects funded by agencies in the north and in the south of Europe, or about the global evolution of funded projects towards the study of more complex questions, focusing more on socio-ecosystems, taking into account human impacts and mixing approaches from several fields of life sciences, instead of focusing on a single biological question. It was also gratifying to to show how TreeCloud could be used as a key component of a broad lexical analysis of a corpus, by designing a protocol to detect topics which were increasingly present in the abstracts, combining TreeCloud with other text analysis tools, especially specificity scores. This was used to contribute to the production of two reports(GELR14; GGE+18), and to give a presentation in Vienna in front of the partners of the project.

With Nadège Lechevrel, a postdoc in the *Biolographes* project supervised by Gisèle Séginger, a professor in French literature at my university, I also had a chance to use TreeCloud extensively to analyze a corpus of articles from the *Revue des Deux Mondes* written by scientists or by men and women of letters, to compare their perspectives on common matters. To this aim, after building a general tree cloud visualization of the whole corpus, we focused on the contexts of the 10 most frequent words. This allowed to give a detailed description of the scientific knowledge of the 19th century which appeared in those articles, and in particular, the aspects which were treated more either by the scientific or of the literary community. We collaborated again in another project supervised by Gisèle Séginger, *Animalhumanité*, also with Tita Kyriacopoulou and Claude Martineau. There, we used the same corpus to investigate the way the word "étude" is used.

With my sister Christelle Mariotte, we also used TreeCloud to analyze the answers she obtained in an online form with open questions about the care of elderly in the Alpes-de-Haute-Provence. I usually use the obtained tree clouds to show that such visualization tools can be used to directly give decision-makers direct access to the actual words of the respondents, without having to include sample verbatim responses that are supposed to reflect the opinions mostly expressed among respondents. Whenever answers are short and similar, it is possible to even guess the main sentences by reading the tree cloud. The one built for the question about what could be improved illustrates how such a visualization can be read, guided by the verbs, to summarize what people think would be important to improve.

The work with the colleagues involved in the ANR project APPEL, supervised by Jean-Gabriel Contamin, a professor in political science in Lille, also involved the use of TreeCloud to analyze words in the beginning (more precisely, among the first ten words) of petitions of the website *lapetition.be*, even if they were not included in the paper accepted at the JADT 2018 conference (BDG$^+$18), by lack of space. Figure 2.1 illustrates a previous finding of Jean-Gabriel Contamin (Con01), the fact that the texts of petitions may target both the decisions makers who will receive it and the people who will sign it. A complementary manual analysis showed that 6.5% of the texts of petitions in this category were explicitly targeting decision makers and 3.8% were explicitly targeting people who would sign the petition.

2.2.2 Making edge lengths more meaningful in TreeCloud

Finding a way to compute more meaningful edge lengths was a project I started during my postdoc with Alain Guénoche in Marseille, nourished by discussions with him and heavily inspired by the article (GG02), in a collaboration with Nuria Gala and Alexis Nasr, who were based in the nearby LIF laboratory (GGN12).

Indeed, the edge lengths computed by the tree reconstruction algorithm, Neighbor Joining (SN87), usually provide a tree whose center is barely readable, as the edge lengths are quite often large for edges adjacent to leaves, and very small for edges deeper inside the tree. As tree clouds are used in practical applications to identify classes of words grouped together into one subtree, it is not always easy to clearly distinguish the limits of subtrees attached deeply inside the tree. Furthermore, the "true" edge lengths of the trees built by the Neigbor Joining algorithm are expected to best reflect the input distance matrix globally, rather than to facilitate this interpretation of the tree as a hierarchy of interrelated word classes.

On the contrary, the nature of the cooccurrence data used to build the distance matrix which is represented by this tree should even preclude from trying to analyze the distances in the tree: it is possible that two words A and B which never cooccur in a text, often cooccur with a third word C. We would therefore expect them to be both located close to C in the tree, which would imply a great discrepancy between their real cooccurrence distance and their distance in the tree. This distortion actually illustrates that the many mathematical formulas evaluating the global proximity of two words, based on their cooccurrence inside a text, have no reason to provide tree-like distances, or, more formally, to output a distance matrix which respects the four-point condition (Bun74).

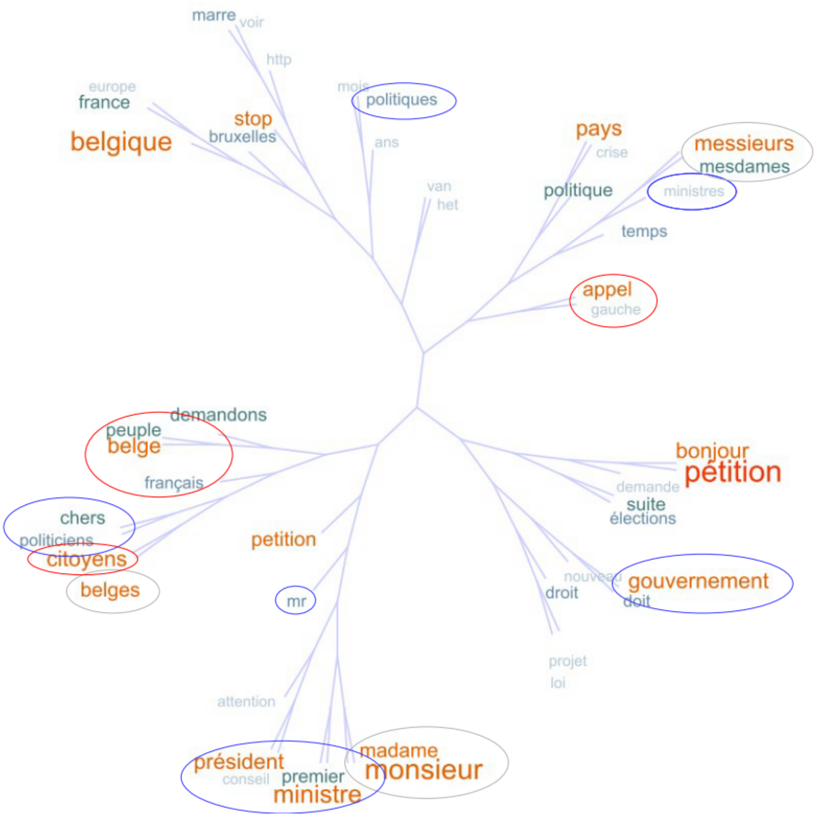


Figure 2.1: Tree cloud of the 50 most frequent words in the beginnings (10 first words) of texts of petitions in the political category of the website lapetition.be: words suggesting that decision makers are targeted by the beginnings of texts of the petition are surrounded in red and words suggesting that people who will sign the petition are targeted are surrounded in blue.

The idea for computing edge lengths which would be more meaningful was therefore to still use the output of the Neighbor Joining method for the topology of the tree, which seemed to provide relevant results in practice, while adding a postprocessing step which would recompute edge lengths which would be well-suited for the goal of visually extracting classes of words from the tree.

We therefore designed a protocol based on trees of words built in two different ways to test the performance, for the task of building semantic partitions of words, of the five formulas[2] provided in (GG02), which evaluate how separated two subtrees are in an unrooted

[2] The first formula, called *computedLength*, is simply the original edge length computed by the Neighbor Joining algorithm. The second one, called *triples*, measures the rate of subsets of 3 leaves $\{a, b, c\}$ such that $d(a, b) \leq \min(d(a, c), d(b, c))$, among those where a and b belong to the same subtree and c to the other one. The third one, *quartets*, measures the rate of subsets of 4 leaves $\{a, b, c, d\}$ such that $d(a, b) + d(x, y) \leq \min(d(a, x)+d(b, y), d(a, y)+d(b, x))$, among those where a and b belong to the same subtree and x and y to the other one. The fourth one, *lengthRatio*, is value obtained by dividing the average distance between words separated by the edge between the two subtrees, by the average distance between words located in the same subtree. Finally, the fifth one, *agreementPairs*, is the value obtained by dividing the sum of the number of pairs of words located in the same subtree which have a distance at most d_m and of the number of pairs of words

phylogenetic tree, based on the distances between pairs of leaves of the same subtree or on two different subtrees. For both parts of the protocol, a partition of words into k subsets of semantically similar words was built manually to be compared with one automatically built in the following way, from a tree of words whose edge lengths were computed with one of the 5 formulas to be tested: the $k - 1$ longest edges were deleted, which split the graph into k connected components, which partitioned the words into at most k subsets (some connected components may not contain a leaf, and therefore would not generate a subset of words in the output partition).

The first set of trees of words was built from the *Polymots* database, where among 20000 words grouped together into 2000 families, each centered on a root word, 20 families were selected to manually partition them into semantic classes. For example, here is an example of the partition for the root word « art » in French: {{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}[3]. Then, for each of these 20 families, a composite distance taking into account word cooccurrence in the TLFI as well as the number of common affixes of the words was computed and used as the input data to build the tree.

The second set of trees of words actually consisted in one tree built in the following way, inspired by an experiment performed by Jean-Marc Leblanc with his tudents: 10 groups of master students were assigned the task of writing a text of more than 300 words which has to contain the 25 given words, trying to put together the words contained in each subset of the partition[4]. The initial partition of the 25 given words was actually built from the tree cloud of a corpus of three newspaper articles in French about Wikileaks[5].

In both cases, two formulas among the five tested, *triples* and *lengthRatio*, provided significanly better results for the corrected Rand score, which computes the agreement, between the manually built partition and the one built automatically using edge lengths computed with this formula.

2.2.3 Maintaining and developing TreeCloud

Maintaining TreeCloud to keep it available since 2010, while trying to develop its functionalities, has not been an easy task. The tool was first available as a desktop program which required the installation of Java, Python and SplitsTree. The latter is a Java program which

separated by the edge between the two subtrees which have distance at least d_m, where m is the number of word pairs located in the same subtree and d_m is the m^{th} smallest value among the distances between pairs of words.

[3] These words could be translated by the following ones in English, using the *Linguee* website: {{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}.

[4] The assignment, given in French, was to write “un texte de plus de 300 mots qui fait obligatoirement apparaître les 25 mots voulus, en tentant de rapprocher les mots contenus dans une même classe de la partition”.

[5] “WikiLeaks : une transparence qui fait débat” was published by Alexandre Piquard in *Le Monde* on November 30th, 2010, “WikiLeaks change la donne de la diplomatie et des médias”, by Nicolas Rauline in *Les Échos* on November 29th, 2010 and “Wikileaks, une nébuleuse si peu transparente…”, by Pierre Demoux in *Les Échos* on December 9th, 2010.

reconstructs phylogenetic trees and networks from various kinds of input data, mainly developed by Daniel Huson. I had contributed to the development of a network visualization optimization in the software during an internship with Daniel Huson in 2005.

The web version was developed mainly by a bachelor student, Jean-Charles Bontemps, who decided to program it as a voluntary side project of the Python courses I was teaching in Montpellier, and to make it available under the GPL license. It consists of an interface in HTML/CSS to enter the text and the parameters and a CGI script in Python which calls a C program to build the tree cloud and output a webpage which displays the output visualization as an SVG file augmented with some Javascript code to make it interactive. A few functionalities are missing in this web version, such as "chronological" coloring, which colors words depending on their average position in the text, or displaying the real edge lengths obtained by the tree reconstruction algorithm instead of unit lengths for all edges of the tree.

I used the desktop version to develop new functionalities such as the computation of edge lengths described in the previous section. Given the large number of software developed in the French community of statistical textual analysis (Alceste, Hyperbase, Lexico, TXM, Textobserver, Cortext, just to name a few) the strategy to develop TreeCloud further, instead of spending a large amount of time and resource on the web version or the desktop version, was to provide implementations in other languages in order to facilitate the inclusion of tree cloud visualizations into other software, and to add new functionalities as prototypes in the desktop version based on SplitsTree.

After my arrival at LIGM in 2011, Patrice Hérault, an engineer at the university, accepted to migrate there the web version previously hosted at my previous lab, LIRMM. Claude Martineau, who was developing lexical resources at LIGM for the Unitex software[6], suggested adding a preprocessing step using Unitex on the web version. This would result in new functionalities to filter the words in the tree clouds, for example keeping only nouns, or removing words based on their part-of-speech label, using the linguistic resources available in different languages in Unitex, integrated by Claude Martineau (Mar17). He also corrected a default of the web version which would remove from the text all words not kept in the tree cloud, instead of taking them into account as word separators in the cooccurrence computation step.

In 2014, when I organized the EPIT summer school with Stéphane Vialette, I received what could have been considered as a spam application from an Indian student from IIT Ropar, Deepak Srinivas. After contacting him to ask whether he was indeed willing to participate in the event, he replied that it was not the case, but he was looking for a summer internship. After complications with his visa, which he obtained at the last minute thanks to the scientific advisor at the French embassy in Chennai, he was able to come to France. He implemented the algorithm by Barthélemy and Luong (BL87) in C++ inside the online version of TreeCloud and also incorporated a new tree visualization library, simply based on a force-directed algorithm (which is relevant if we do not care about edge lengths), to display the result.

[6] Unitex is a collection of programs for text analysis using linguistic resources, based in particular on the use of local grammars

In 2016, Cristian Martinez, a doctoral candidate with Tita Kyriacopoulou at LIGM, who was improving Unitex as well as setting up best practice for its continuous development, suggested applying with Unitex to the Google Summer of Code program, which supports the development of free software by funding student internships. The application he prepared for Unitex was successful and in the Summer 2016, I was able to supervise a Russian student, Aleksandra Chaschina, to develop a Java version of TreeCloud, which could be integrated into the Unitex/Gramlab software. Her code was also integrated into the TextObserver software by Yacine Ouchene, who was hired as a software engineer by Jean-Marc Leblanc, with ANR APPEL funding.

Aleksandra Chaschina was also involved in the Javascript version of Tree Cloud. In 2015 I had met Yu Zheng at NUS in Singapore, who was able to quickly code, using D3, the equal-angle algorithm for phylogenetic tree visualization[7]. Collaborating with the BiodivERsa program mentioned above, provided the opportunity to fund a short work contract for Aleksandra Chaschina when she arrived in France for her studies at Université Paris Sud. Using the code by Yu Zhen to visualize the trees, she developed a web interface to explore data using tree clouds as well as specificity scores[8]. Unfortunately, there was not enough time to properly test the computation of specificities, which is currently not accurate, but the tree cloud building part works well, and should be made available as an independent easy-to-use Javascript library.

We tried to summarize all this with Teresa Gomez-Diaz when writing a software management plan for TreeCloud, which illustrated the difficulties of trying to develop versions in several programming languages of a text visualization method, with limited resources.

2.3 Detecting temporal proximity in text documents

As seen in the previous chapter, hierarchical clustering can be used in digital humanities to analyze texts. It is also widely used to classify texts depending on their content. When I started my six-month leave[9] in September 2020 at Lattice in Montrouge, I talked with Olga Seminck, who had planned to use hierarchical clustering as a first step, and a regression analysis, as a second step, to analyze the evolutions of the idiolect of several 19th century novel writers. The results of our work, also conducted with Thierry Poibeau who supervised the project, and with Dominique Legallois, were published in (SGLP22), but along the way, we proposed new ways of addressing the issue of detecting a chronological signal in

[7] https://github.com/adamzy/PhyloPlot/

[8] https://treecloud.univ-mlv.fr/treecloud-corpus/voeux/

[9] I did not really understand why the head of our lab, Stéphane Vialette, insisted that we should take a *délégation CNRS*, a semester or year fully dedicated to research, with no teaching, until I was finally convinced in 2019 to apply for one in 2020. Although the moment probably wasn't the best, with the second COVID-19 lockdown in France, I just had enough time to enjoy meeting the colleagues at Lattice in September and attending three days of training about handling data in research projects in the humanities and in social sciences organized by the CNRS in Aussois, before having to continue working from home. This time was however very important to have the feeling that I was fully transitioning to research in digital humanities, and most of all to finally take the time to write and publish on the MPRI website an offer for a master internship, which would lead to the results described in section 3.3.

a distance matrix of dated elements, or in a tree built from this matrix, therefore evaluating the relationships between content similarity and temporal proximity.

2.3.1 Testing a matrix for the chronological signal

A pioneering work about linking similarity data with a chronological signal had been done by Robinson in (Rob51), in the context of archeology, where the goal was to reorder pottery fragments chronologically based on their similarities, making the assumption, sometimes called the rectilinear hypothesis, that the more distant apart in time fragments are, the less similar they should be. A dissimilarity d between a set X of elements is therefore said to be *Robinsonian* if the elements of X can be ordered such that $\forall i < j < k \in X$, $max(d(T_i, T_j), d(T_j, T_k)) \leq d(T_i, T_k)$. This definition is illustrated in figure 2.2.

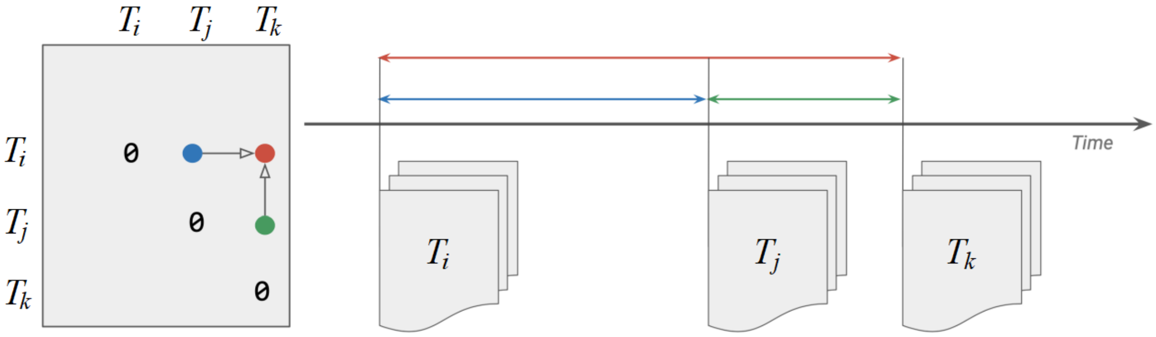


Figure 2.2: An illustration of the property defining of Robinsonian dissimilarities d: for all $T_i < T_j < T_k$, $max(d(T_i, T_j), d(T_j, T_k)) \leq d(T_i, T_k)$.

We therefore defined the Robinsonian score of a dissimilarity d as the proportion, among triplets of elements ordered chronologically, of those which respect this condition. Comparing this score with the ones obtained by randomly reordering the elements would allow us to estimate a p-value for the observed Robinsonian score, to evaluate whether the value of Robinsonian score would be too high to be obtained by pure chance.

The computation of this Robinsonian score was implemented by Olga Seminck and tested on the CIDRE corpus of novels by 11 writers of the 19th century, in an experiment described in section 4 of (SGLP22) using textual features called *motifs* (LCL18) to compute the dissimilarities between the texts, using formulas implemented in the R package called *stylo* (ERK16). It showed that for all authors except Sophie de Ségur, the Robinsonian score of the dissimilarity of their dated works is higher than what could be observed in more than 98% of 10 000 random orderings of their works, which is therefore considered to be statistically significant.

2.3.2 Testing a tree for the chronological signal

Before coming up with the Robinsonian score idea, preliminary tests prepared by Olga Seminck consisted in using the R package stylo to create dendrograms, that is to say rooted clustering trees to visually check whether texts whose publication dates were similar appeared close to each other in the tree.

Although such an approach is frequent in statistical textual analysis, as can be seen with the examples of trees given in figure 2.3), the comments of such trees are often limited to noting that the tree structure obviously reflects a chronological organization, without evaluating it more precisely.

As can be seen on figure 2.3, reordering the children of the second child of the root, by putting its first child leading to the 1978 presidential address between its third child (above all addresses of years 1969 to 1973) and its fourth one (above all addresses by François Mitterrand, of years 1981 to 1994), would improve the visual perception of this chronological signal in the tree. However it is not possible to obtain the perfect chronological order on the leaves by simply reordering children in the tree, as can be seen for example with this 1978 presidential addresses, which cannot be placed between the ones of 1977 and 1979 simply by child order changes.

Therefore, in order to investigate the question of evaluating whether a clustering tree displays the chronological signal of the elements it classifies, the relevant problems would be (GSLP21):

- Is it possible to reorder children in the tree so that the resulting order of the leaves matches the chronological order perfectly?
- If this is not the case, would it be possible to reorder them so as to optimize consistency between the order of the leaves and the chronological order?
- Would it be possible to obtain by pure chance such a level of consistency between the optimal order of the leaves and the chronological order?

The first two questions reminded me of tanglegram problems, which I was teaching in my courses about algorithmics for bioinformatics in the first year of master in computer science at ENS Paris-Saclay. The most classical for the *tanglegram* problem consists, for two binary phylogenetic trees T_1 and T_2 having the same set of leaves, in deciding whether it is possible to reorder the children of its internal nodes so as to make the orders of their leaves as similar as possible. The notion of similarity is counted, here, as the number of conflicts between the two orders, that is the number of pairs of leaves which are not ordered in the same way in both orders, which can also be viewed as the number of inversions in the permutation obtained by labeling the leaves of the first input tree by 1 to n from left to right and reading the order of labels on the second tree, starting from the same side of the tanglegram.

The problem, which is NP-complete, is linear time solvable if the number of expected conflicts is 0 (it can easily be turned into a planar graph detection problem on the graph obtained by connecting by edges the roots of the trees and the leaves having the same label), and solvable in $O(n \log n)$ if the order of leaves of the first tree is fixed (VASJG10), which corresponds exactly to our second question above if the clustering tree is binary and if the dates of texts are all distinct, which allows to fix the order of the first input tree accordingly.

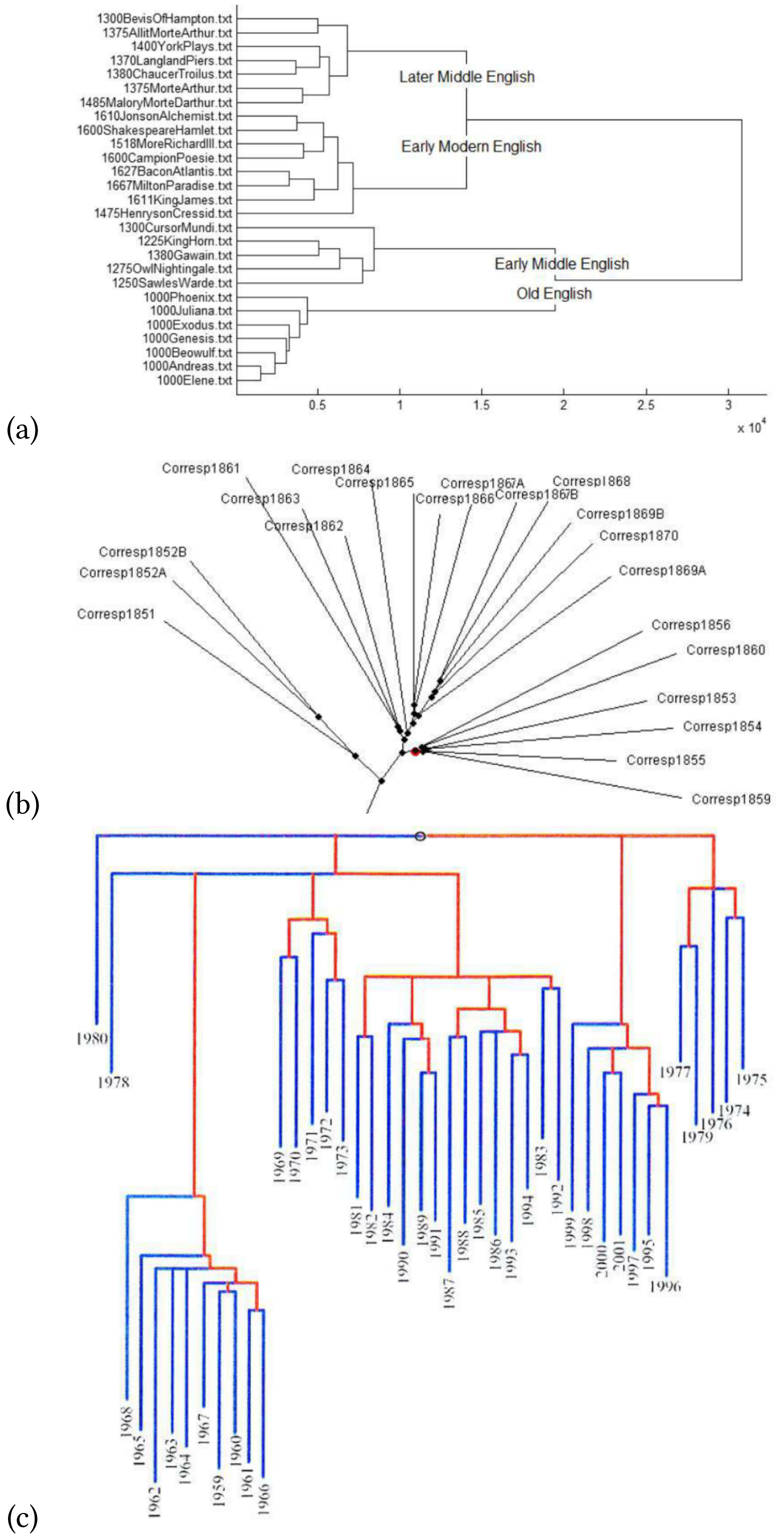


Figure 2.3: Hierarchical clustering trees of old English texts (figure extracted from (Moi20)), letters written annually by Victor Hugo (figure extracted from (LL13)) and new year's presidential addresses in France (figure extracted from (Leb16)).

The problem, taking as input a tree T with labeled leaves and a strict order on the leaves of T, is then called *OTCM* for One-Tree Crossing Minimization. I therefore implemented in Python a more simple $O(n^2)$ version of this algorithm, also described in the same article[10].

But the complexity of the problem in the general case of non-binary trees was not clear and we called Laurent Bulteau to the rescue to study it. His intuitions quickly resulted in a proof of NP-completeness for the problem, as well as for several variants (BGS22).

One of the variants which is interesting in terms of applications is *OTDE*, One-Tree Drawing by Deleting Edges. In this variant, the optimization criterion to minimize is the number of leaves to delete so that all conflicts disappear. This optimal value corresponds to the permutation having the longest increasing sequence, which allowed us to design a fixed parameter algorithm, parameterized by the maximum degree of the tree, based on a dynamic programming approach inspired by the dynamic programming algorithm solving the search for the longest increasing subsequence in a permutation. I also implemented this algorithm, which runs in $O(d!n^{d+2})$, where d is the maximum degree of the tree, in Python[11]. An illustration of the output of this algorithm is shown in figure 2.4.

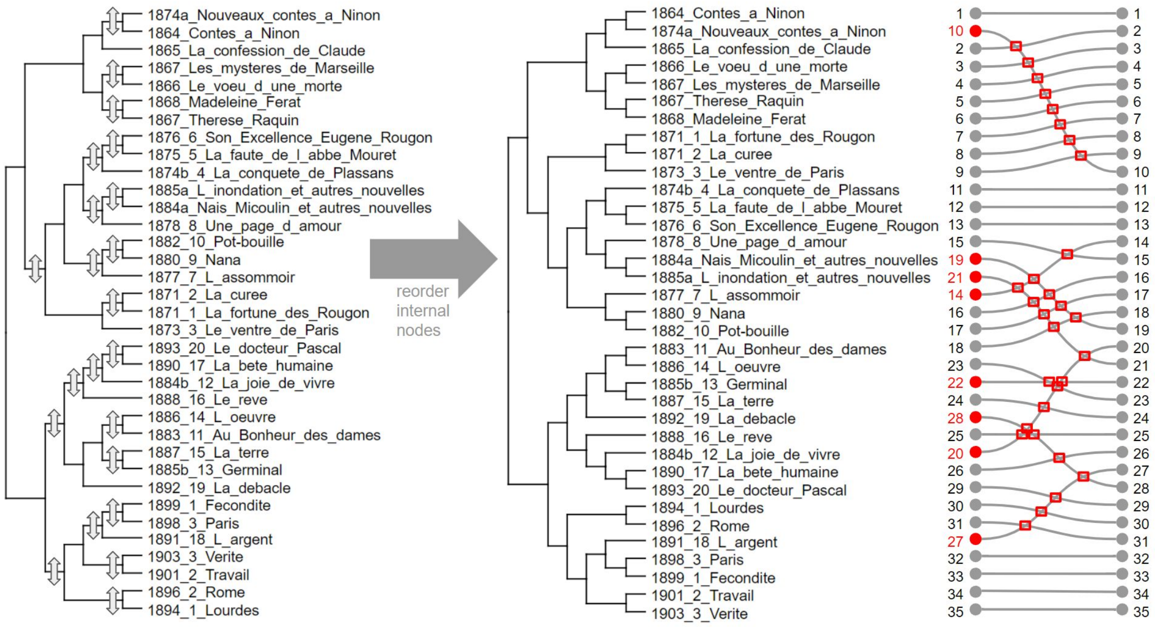


Figure 2.4: An illustration of output of the script solving *OTCM* and *OTDE* on a tree of novels published by Émile Zola. Children of the internal nodes were reordered as shown by vertical arrows to obtain an order of the leaves which has 30 conflicts with the chronological order (shown with red boxes) and where 8 leaves can be deleted (the ones colored red) so that the remaining leaves are in chronological order.

[10] The implementation is available at https://github.com/oseminck/tree_order_evaluation.

[11] Available at https://github.com/oseminck/tree_order_evaluation.

2.3.3 Analyzing idiolect evolution of XIX$^{\text{th}}$ century authors

The two methods described in the previous sections were applied to a corpus of novels by several prolific writers of the XIX$^{\text{th}}$ century (SGLP22).

I contributed to the preparation of the CIDRE corpus (SGLP21a; SGLP21b) by providing scripts to automatically download novels in the ePub format from various sources, using simple HTTP requests for Project Gutenberg, BeQ and Gallica, and web browser automation with the selenium library for Wikisource[12]. Those scripts were later improved by Olga Seminck[13].

A key issue to build this corpus of more than 400 fiction books by a total of 11 authors was to estimate the writing date of the novels. I helped to establish this data, automatically by parsing the chronologically sorted results of queries to the general catalogue of BnF, the French national library, to find the date of the first edition, sometimes manually by reading the author's page on Wikisource (where the first publication date is often provided, even when the novel was first published as a serial in a newspaper) or by reading the preface. Indeed, in the case of George Sand, for example, she who would often give context about writing the novel, including the date, in the preface of new editions.

The methods presented in the previous sections allowed to identify authors for whom there was a clear temporal signal in the textual data. A regression method was also used to learn, for each author, a model to date their works based on part of their production and guess the date of their remaining novels. This evolution of idiolect was then analyzed for four authors for which good results were obtained, Honoré de Balzac, Daniel-Lesueur, George Sand and Émile Zola (SGLP22), using the motifs highlighted by the regression analysis as the most useful for the automatic estimation of dates.

I focused on Daniel-Lesueur, where the following motifs are globally increasing over time:

- a dash (for dialogue) followed by "quel", "quelle" or "quels" ("which"), followed by a common noun, followed by a question mark;
- a dot (end of the previous sentence), followed by an adverb, followed by a comma;
- "d'un tel" followed by a common noun (possibly in the feminine or plural form).

On the contrary, the following motifs are decreasing:

- "dire" ("say"), followed by a pronoun ("he", "she", etc.);
- "dans lequel/laquelle/lesquels/lesquelles" ("in which");

[12] I thank François Briatte for the discovery of selenium, this is one of the many examples where people I have met in the *Confédération des Jeunes Chercheurs* have made me discover new tools or practices which have been helpful in my research.

[13] Scripts step1-getEBooks.py and step2-convertToTei.py available at https://github.com/oseminck/cidre/tree/main/scripts.

- “d’un tel” followed by a common noun (possibly in the feminine or plural form).

The first increasing and decreasing motifs mentioned above suggest that Daniel-Lesueur changed her way of writing dialogues, choosing a more concise and energetic style, which seems to be consistent with the evolution of the average sentence length, globally decreasing in her works.

It is also interesting to see that distant reading tools such as Lexico (LMF+02) confirm some of the observed phenomenons on such higlighted motifs, when computing the over or underrepresentation of some words in novels sorted by date (using the “specificity score” of Lexico). For example, the expression “dans lequel/laquelle/lesquels/lesquelles” (“in which”) is statistically overrepresented in several early novels and underrepresented in later ones, but the shorter equivalent “où” (“where”), is, on the contrary, underrepresented in earlier novels and overrepresented in later ones.

2.4 Exploring local heritage with digital applications

In 2017, the selection of the I-Site Future project for funding by the Programme d’Investissement d’Avenir, run by the French national research agency, ANR, resulted in local incentives to encourage research on sustainable cities, which were confirmed after the creation of Université Gustave Eiffel in 2020 and the confirmation of the I-Site Future label of the university in 2022.

This created oportunities to start interdisciplinary projects with colleagues from other laboratories of Université Gustave Eiffel, where the topics of historical heredity and urban proximity would play a major role.

2.4.1 Uncovering the industrial past of cities

Involved in the Cité des dames project (see next section), whose goals included automatically identifying urban locations in texts, Catherine Dominguès contacted me as she was looking for funding for a doctoral project on automatically extracting information about areas polluted by industrial activities from various kinds of administrative documents. I gladly accepted to co-supervise the project with her and after she managed to secure funding for the project (half of the funding from ADEME, the French agency for ecological transition, and the other half from IGN, the national institute of geographic and forest information), we hired Chuanming Dong who started his doctorate on this topic in October 2019.

A first challenge of this doctorate was to get textual data from several databases about areas which may have undergone pollution or which are currently hosting industrial activities which may have a negative impact on the environment. As several databases existed on this topic in the French administration, for example Aria, BASOL and BASIAS, an important task was to identify identical loctions in those database, based on the proximity of these locations and the similarity of the rest of their metadata (Don23).

Then, Chuanming extracted a corpus of texts from BASOL and worked on the automatic extraction of events from these documents, for example pollution events or administrative events about how to deal with this pollution, or about the company itself. The French language model CamemBERT, created in 2019 (MMOS^{+}20), was used to transform words of the text into mathematical vectors, and deep learning methods (more precisely, Bi-LSTM neural networks) were developed by Chuanming to look for various informations about those events in the text (DGD21). He also developed a hybrid method, including semi-automatic unsupervised learning methods and automatic supervised learning method in order to design a semantic classification of these events (DGD22). He also co-supervised two students project to build a database and a map to display the extracted information about the areas linked with industrial pollution. Together, we supervised the internship of one of my first-year students, Noam Sebahoun, to obtain a practical web interface to explore the extracted data, which was later improved during the internship of another student, Clément Sicot, supervised by Catherine Dominguès[14]. Finally, this proof-of-concept tool shows how we can use it to focus on an area and uncover its industrial past, with a possible heritage of pollution to be dealt with by future generations.

2.4.2 Designing guided tours to discover women's heritage in cities

It was during working sessions with Caroline Trotot in 2016/2017, for the ECLAVIT project[15], that I realized that women authors were often underrepresented in the literature corpora I had been studying in previous research projects, in class during high school, and on my own bookshelves. Further discussions on this topic resulted in my participation in the #HackÉgalitéFH hackathon, a one week-end event in March 2017 where, in a team of 5 people, we developed the prototype of a website to encourage literature teachers to study more texts written by women with their students. The project was awarded one of the three prizes of the event, presented by Laurence Rossignol, the minister of women's rights. It also led to the creation of the association *Le deuxième texte* in September 2017, and to the VisiAutrices research project[16], which, in particular, produced a collection of digital books in the public domain written by women[17]

With Caroline Trotot, we thought that this collection of texts could be one of the ways to make the cultural heritage of women more visible in cities. This was one of the motivations

[14] https://clementsct.github.io/Carto_IGN_LIGM/

[15] The ECLAVIT project was supervised by Tita Kyriacopoulou and aimed at making several digital applications for text mining or text analysis developed in Université Paris-Est more interoperable (https://eclavit.hypotheses.org/).

[16] This project I directed from 2017 to 2019 was attributed a funding of 18218 euros by the CNRS and the RnMSH (National network of Houses for the Social Sciences and Humanities). It resulted in particular in the *Histoires d'autrices* website, available at https://citedesdames.github.io/histoires-autrices/, where several datasets about literature teached or published in France can be explored to observe the evolution of the percentage of women among authors, and get more information about the canonicity of these women writers (see e. g. (BCP23) which used the data for a quantitative analysis of canonicity in French 19th and 20th century literature).

[17] Part of this collection can be explored at https://treecloud.univ-mlv.fr/philologic/visiautrices within the Philologic web interface, developed by the University of Chicago and the ARTFL.

to start the *Cité des dames, créatrices dans la cité* project, which was funded by Université Gustave Eiffel from 2019 to 2023. One of the goals of this project was to provide tools to help building guided tours of cities to discover how they were shaped by the creations of women from the past. Automatically detecting urban locations in books written by women was one of the ways to enrich those guided tours, and connect some places in cities with the women who wrote about them. This motivated in particular the study described in Section 3.2.2 to improve the results of named entity recognition algorithm, in particular for geographical named entities, in texts of the 16th or 17th centuries.

This project also resulted in the development of several digital tools to discover the cultural heritage of women in some French cities, which were mentioned in a Télérama article in February 2024 (Fau14). In 2019, I supervized the internship of Gilles Avraam, who coded the game Matrimoine GO ![18], based on an idea by Edith Vallée, the author of a book describing 20 guided tours, one for each Paris district, of several topics about the cultural heritage of women (Val18).

In 2021, I supervized the internship of Ulysse Gravier, who developed a web application for guided tours, *Les promenades du matrimoine*[19], designed for smartphones, illustrated in figure 2.5. I contributed to the creation of the *Promenade des Marguerite*, a guided tour in Paris about writers named Marguerite, with some colleagues from LISAA (a variant of this guided tour was prepared, using the content we had gathered, by Sonia Thuillier, a tour guide for *Feminists in the City*, during the final workshop of the *Cité des dames* project in June 2023.). I was also involved in the creation of the guided tour of Marie de Gournay, a 17th philosopher, in Paris, with Suzanne Duval and her intern, Maéva de Sousa. I also added extra content to a guided tour about Marceline Desbordes-Valmore in Douai designed by several administrations of the city who had collaborated with members of the Société des études Marceline Desbordes-Valmore in 2021 for the Festival Résonances. In August 2023, I was able to experiment it by guiding a small group of teachers of French literature in Douai, in a visit which was co-organized with the municipal library of Douai and the Musée de la Chartreuse, who also offered a visit of their collections.

The same year, I also supervized the internship of Alexis Martinet who coded a web application to explore historic travel narratives, where text or multimedia documents are associated with each visited city mentioned in the text or in archives about the travel. We used it, with Nicole Dufournaud and Caroline Trotot, for a study of Marguerite de Valois's journey to Flanders in 1577, described in her *Memoirs*[20], as well as in a collaboration also with Caroline zum Kolk about the grand Tour de France of Charles IX and Catherine de Médicis from 1564 to 1566[21] and in a collaboration with Fanny Boutinet to represent Catherine de La Guette's journey to Bordeaux during the Fronde[22]. This application and its use

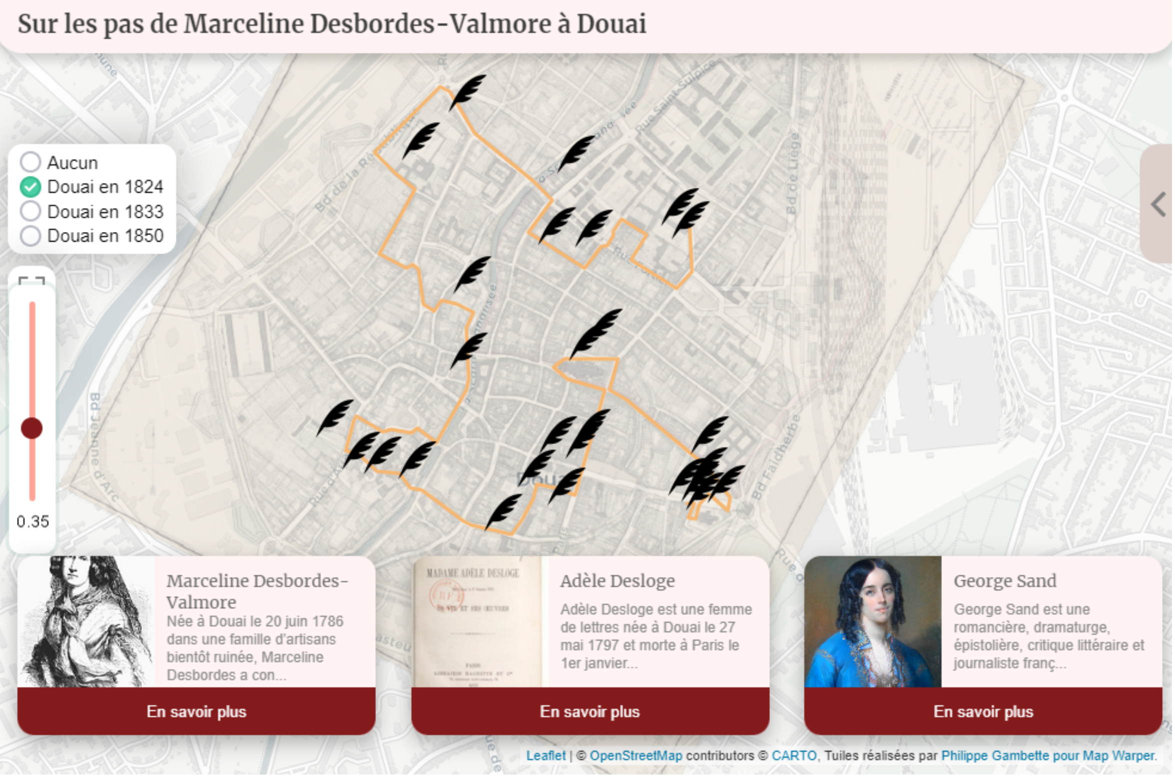
[18] https://citedesdames.hypotheses.org/110

[19] https://ulysseee.github.io/les-promenades-du-matrimoine

[20] https://citedesdames.github.io/de-ville-en-ville/?site=1. A presentation of this journey and the application was given in a talk by Nicole Dufournaud, Caroline Trotot and myself, entitled “Identifier, représenter, penser les réseaux urbains de l'humanisme autour de Marguerite de Valois : l'exemple du voyage des Flandres (1577)” at the workshop *Les femmes dans les réseaux urbains de l'humanisme (1492-1615)* in Bordeaux on June 30th, 2022.

[21] https://citedesdames.github.io/de-ville-en-ville/?site=0

[22] https://citedesdames.github.io/de-ville-en-ville/?site=3



Sur les pas de Marceline Desbordes-Valmore à Douai

Marceline Desbordes-Valmore
Née à Douai le 20 juin 1786 dans une famille d'artisans bientôt ruinée, Marceline Desbordes a con...

Adèle Desloge
Adèle Desloge est une femme de lettres née à Douai le 27 mai 1797 et morte à Paris le 1er janvier...

George Sand
George Sand est une romancière, dramaturge, épistolière, critique littéraire et journaliste franç...

Leaflet | © OpenStreetMap contributors © CARTO, Tuiles réalisées par Philippe Gambette pour Map Warper.

Figure 2.5: An illustration of the web application *Les promenades du matrimoine* with the guided tour about Marceline Desbordes-Valmore in Douai.

in research, teaching and scientific mediation were presented in a talk by Fanny Boutinet, Caroline Trotot and myself, also prepared with Nicole Dufournaud, entitled “Visualiser les itinéraires historiques de récits de voyages féminins”, at the workshop *Visualiser les récits de voyage à l'ère du numérique : approche interdisciplinaires*, on May 31th, 2024.

In 2022, I supervized the internship of Danie Lancea who implemented the website *Empreintes de femmes*[23] in order to display data on more than 100 guided tours about the cultural heritage of women in France. This website, as well as the associated database, is useful to have an overview of the diversity of such guided tours, called *promenades du matrimoine*, which have increasingly been developed since 2015.

Those guided tours sometimes mention women appearing on street names or building names. I also coded two web applications on this topic. *Plaques du matrimoine*, which I started in 2019 to visualize streets or buildings named after women in a city, using data from OpenStreetMap and from Wikidata, was improved in 2020 by an intern, Alan Akra[24]. With *Nom d'une plaque*, one can search for the name of a person and display on a map of France all streets named after her or him, using data from the *Base Adresse Nationale*, which can also be exported as a table where the naming year is provided, extracted from an export of the FANTOIR database[25].

[23] https://citedesdames.github.io/empreintesdefemmes/
[24] https://perso-etudiant.u-pem.fr/~gambette/PlaquesDuMatrimoine
[25] https://citedesdames.github.io/NomDUnePlaque/

This research project was a great opportunity for interdisciplinary collaborations, both with colleagues in literature, and with Nicole Dufournaud, a specialist in digital humanities and in women's history. In particular, our cosupervision of interns, some of them in literature such as Beatrice Mundo who wrote her Master's Thesis about family and power in the correspondence of Catherine de Médicis, under the supervision of Caroline Trotot, was also a good way to work together.

Chapter 3

Uncovering Heredity from Similarity in Digital Humanities

The alignment of a set of homologous genes or proteins among different species is a classical first step before reconstructing their evolution history. A similar approach can be performed in digital humanities when studying the different manuscripts of some written work, in order to reconstruct a *stemma* showing the history of copies written by different copyists. In the context of genetic editing, a similar approach can be used to reconstruct the history of different versions written by the author of some published works. Text alignment algorithms can also be used without any genetical reconstruction purposes, for example to identify relevant variants among several versions of a text, or, on a larger scale, to study the historical evolution of spelling, in a language, when aligning the original and modernized versions of large corpora of texts.

In this chapter, I show a few applications of the automatic detection of similarity between texts, using different kinds of text alignment algorithms, in digital humanities, and more precisely in digital edition. First, I will show a few digital tools which were developed for text genetics, as well as for finding possible sources of a text, when editing it. I will then focus on a different kind of heritage with tools developed to analyze the evolution of the French language and to automatically modernize texts, aiming at making them more accessible to the reader today, or to treatment by digital tools. Finally, I will focus on the case of theater plays and new ways of evaluating similarity between plays to uncover influences between their authors.

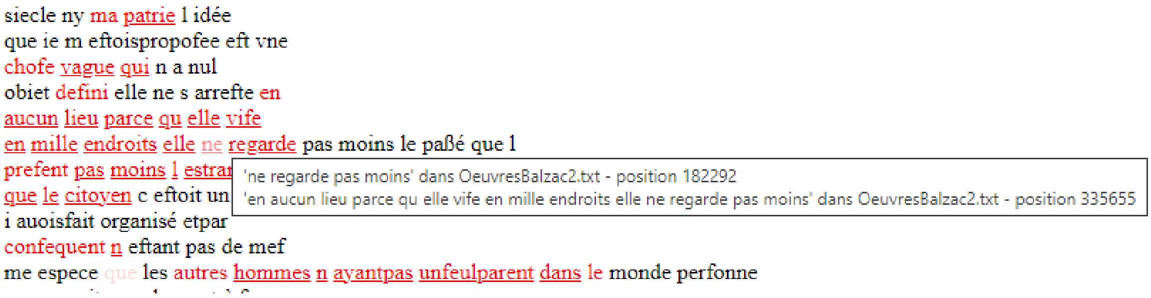
3.1 Text alignment and intertextuality

3.1.1 Automatic search for intertextuality

In 2016, I started to develop a Python script, called intertextFinder[1], to look for intertextuality between a target text and a corpus of possible source texts, simply based on the search for identical 4-grams, i. e. sequences of 4 consecutive words, ignoring punctuation and case, possibly extended if a longer match is detected.

This script was used in particular to look for common hemistichs between the play *Le Favori* by Marie-Catherine de Villedieu and the database of theater plays available online at https://www.theatre-classique.fr/, in a collaboration with Delphine Amstutz, when she worked on a scholarly edition of the play, published by Hermann in 2017. The script also helped Delphine Amstutz to uncover, in 2024, the sources of pages 119 to 145 of Augustin Courbé's 1648 edition of *Le Barbon* by Jean-Louis Guez de Balzac, by looking for them inside Thomas Jolly's 1665 edition of *Les Œuvres de monsieur de Balzac, divisées en deux tomes.* As she suspected, it was a cento of Balzac's correspondence published in the first volume of the 1665 edition as well as 3 other sources found in the second volume, which had not all been identified by previous work. Note that this experiment was performed directly on the text downloaded from Google Books, obtained by optical character recognition, without manual postprocessing: the quality of the OCR is now good enough, even on 17th century texts, so that portions of text of a few paragraphs can be efficiently spotted with such basic techniques as identical 4-gram search.

[1] https://github.com/PhilippeGambette/txtCompare



```
siecle ny ma patrie l idée
que ie m eftoispropofee eft vne
chofe vague qui n a nul
obiet defini elle ne s arrefte en
aucun lieu parce qu elle vife
en mille endroits elle ne regarde pas moins le paßé que l
prefent pas moins l estra  'ne regarde pas moins' dans OeuvresBalzac2.txt - position 182292
que le citoyen c eftoit un 'en aucun lieu parce qu elle vife en mille endroits elle ne regarde pas moins' dans OeuvresBalzac2.txt - position 335655
i auoisfait organisé etpar
confequent n eftant pas de mef
me espece que les autres hommes n ayantpas unfeulparent dans le monde perfonne
```

Figure 3.1: Example of output of intertextFinder, on Augustin Courbé's 1648 edition of *Le Barbon* by Jean-Louis Guez de Balzac, looking for identical 4-grams with Thomas Jolly's 1665 edition of *Les Œuvres de monsieur de Balzac, divisées en deux tomes*, when hovering over the word "ne" before "regarde pas moins".

Inspired by plagiarism detection tools, the output of the script, illustrated in figure 3.1 is a webpage containing the text, where the first word of each 4-gram of the text which was found in the input corpus is colored red, but with a paler color if the 4-gram is found with a high frequency in the corpus: this may indicate that finding this 4-gram in the corpus is actually not significant. Hovering over those colored sequences of words displays a tooltip which lists the texts where it was found, also providing the position of the occurrences in the texts.

I have used the script myself for another application, to quickly identify Marceline Desbordes-Valmore poem's contained in anthologies of her poetic works. This helped to semi-automatically feed the database of more than 6800 editions of her poems I'm maintaining online on the website of the Société des études Marceline Desbordes-Valmore[2], which I have been developing since 2020 and which was open to the public in 2023. The database also references more than 1000 translations of poems in a total of 25 languages. A database of musical score inspired by them poems was also started by Pierre Girod and Françoise Masset, who had gathered about 100 digital scores. Using several sources (including the general catalogue of the BnF, melodiefrancaise.com, RISM, Lieder.net and (Bod86)) more than 600 musical scores are currently available in the database, including more than 400 with a link to a digital version available. Both corpora (editions of poems, and musical scores of poems) will be useful for future projects I am currently planning, about the development of text or text-music alignment algorithms and useful visualizations of their output, in order to facilitate their analysis.

3.1.2 Multi-level text alignment for text genetics or visualization of variants

I have already worked on text alignment algorithms, in particular in the context of the *Cité des dames* project, when wondering about the differences between several versions of texts written by women, for example, to select the most relevant as a reference digital edition for example.

[2] https://societedesetudesmarcelinedesbordesvalmore.fr/oeuvrepoetique/

The online version of the MEDITE tool (GFL04; SGB15), available at http://obvil.lip6.fr/medite/, generally provides very relevant alignments of texts, even being able to spot transfers of sequences of words. However, this software reaches its limits on long texts (comparing distinct editions of books which are several hundreds pages long typically runs for a long time), on cases where more than two versions of the text have to be aligned or in cases where it would be convenient to ignore some minor modifications between the two texts, such as modifications resulting from modernization.

I first coded a few scripts to get around some of these problems. For example, the Python script pairwiseMedite.py[3] simply calls MEDITE multiple times on all possible pairs of texts in the input. It can also be called with different parameters in order to compare successive pairs of portions of the two texts, to reduce the computation time (but this requires to first identify corresponding portions of the texts, such as chapters).

In 2023, I supervised the internship of Maxime Kremer who developed a text alignment web application, COATL[4], in Javascript, in order to run it directly in a web browser (which makes it easier to integrate it into other websites) and to easily export the obtained result. He focused on the case of word substitution in the alignment. The color of a word then depends on how similar it is to the aligned word in the other version of the text: there is a bright purple backgroung if both words are completely distinct, but the color is paler if the words share the same grammatical function or even paler if they have the same lemmatized version, as shown in figure 3.2.

I also focused on the case of order change in editions of collections of texts such as poems or short stories. Using Sankey diagrams to represent such changes of orders was actually an idea already developed by Martin Paul Eve in 2016 (Eve16), when I coded it indepently to represent the changes in the order of two editions of the *Heptaméron* by Marguerite de Navarre, published in 1558 by Boaistuau and in 1559 by Gruget (see figure 3.3), in the script called SankeyCompare[5].

This idea was extended to take into account the case of comparing 3 collections of texts or more, as shown on figure 3.4 by a visualisation of the order of poems in the poetry books published by Marceline Desbordes-Valmore between 1819 and 1830[6].

This visualization allows to quickly identify poems which disappeared in the next edition of the poems, which ones appeared, and which ones changed color, meaning that they are placed in different sections of distinct editions of the poetry books. For example, "Le sommeil de Julien" was first placed in the section of "Romances" in 1819, then placed among the "Elegies" in 1820, then again among the "Romances" in 1822 and 1830.

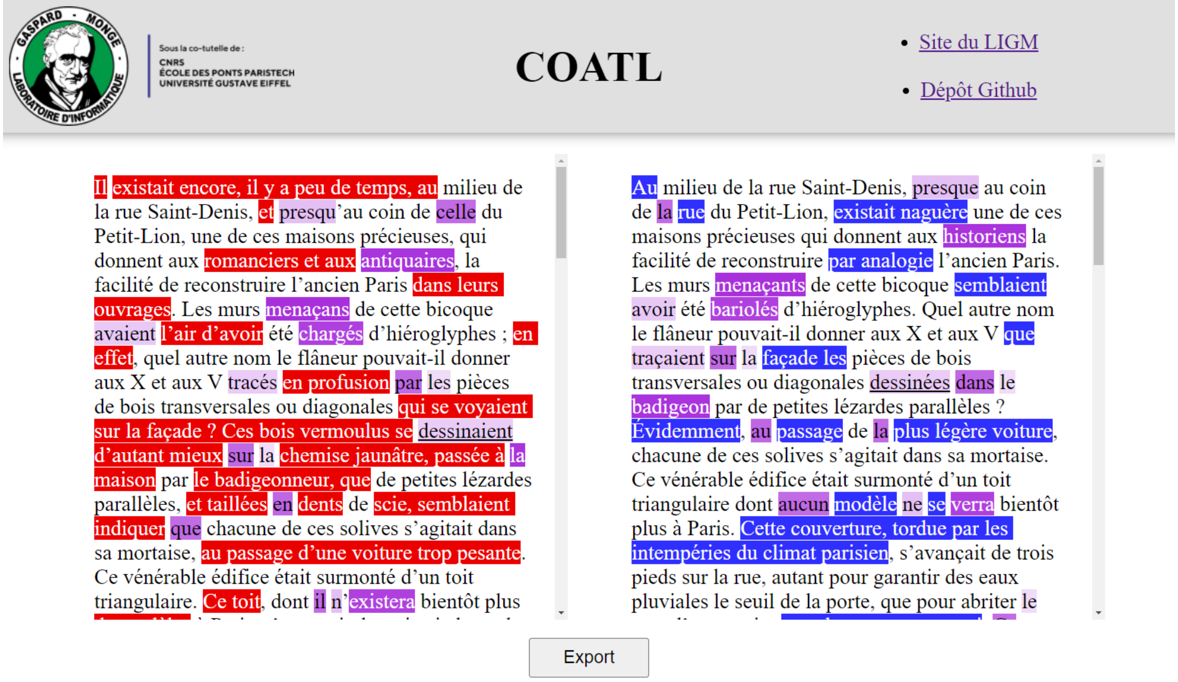
The first feedback from researchers involved in the team supervised by Christine Planté and Andrea Schellino to edit the poetic works of Marceline Desbordes-Valmore was that this figure is not easy to read and that a classical concordance table would be useful. I therefore added a concordance table below this visualisation, making it sortable by any column to answer the needs of colleagues in literature.

[3] https://github.com/PhilippeGambette/txtCompare/tree/master/pairwiseMedite

[4] https://maxb9f.github.io/COATL-LIGM/

[5] https://philippegambette.github.io/txtCompare/sankeyCompare/index.html?id=0

[6] https://igm.univ-mlv.fr/~gambette/2018Visiautrices/MarcelineDesbordesValmore/RecueilsPoesies/


GASPARD MONGE LABORATOIRE D'INFORMATIQUE
Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL
COATL
Site du LIGM
Dépôt Github
Export

Il existait encore, il y a peu de temps, au milieu de la rue Saint-Denis, et presqu'au coin de celle du Petit-Lion, une de ces maisons précieuses, qui donnent aux romanciers et aux antiquaires, la facilité de reconstruire l'ancien Paris dans leurs ouvrages. Les murs menaçans de cette bicoque avaient l'air d'avoir été chargés d'hiéroglyphes ; en effet, quel autre nom le flâneur pouvait-il donner aux X et aux V tracés en profusion par les pièces de bois transversales ou diagonales qui se voyaient sur la façade ? Ces bois vermoulus se dessinaient d'autant mieux sur la chemise jaunâtre, passée à la maison par le badigeonneur, que de petites lézardes parallèles, et taillées en dents de scie, semblaient indiquer que chacune de ces solives s'agitait dans sa mortaise, au passage d'une voiture trop pesante. Ce vénérable édifice était surmonté d'un toit triangulaire. Ce toit, dont il n'existera bientôt plus

Au milieu de la rue Saint-Denis, presque au coin de la rue du Petit-Lion, existait naguère une de ces maisons précieuses qui donnent aux historiens la facilité de reconstruire par analogie l'ancien Paris. Les murs menaçants de cette bicoque semblaient avoir été bariolés d'hiéroglyphes. Quel autre nom le flâneur pouvait-il donner aux X et aux V que traçaient sur la façade les pièces de bois transversales ou diagonales dessinées dans le badigeon par de petites lézardes parallèles ? Évidemment, au passage de la plus légère voiture, chacune de ces solives s'agitait dans sa mortaise. Ce vénérable édifice était surmonté d'un toit triangulaire dont aucun modèle ne se verra bientôt plus à Paris. Cette couverture, tordue par les intempéries du climat parisien, s'avançait de trois pieds sur la rue, autant pour garantir des eaux pluviales le seuil de la porte, que pour abriter le

Figure 3.2: Example of the output alignment computed by COATL for the 1830 and 1855 editions of *La Maison du Chat-qui-pelote* by Honoré de Bazac. Deletions are shown in red, insertions in blue and substitutions in purple, with a paler color if some level of similarity is observed between the two words.

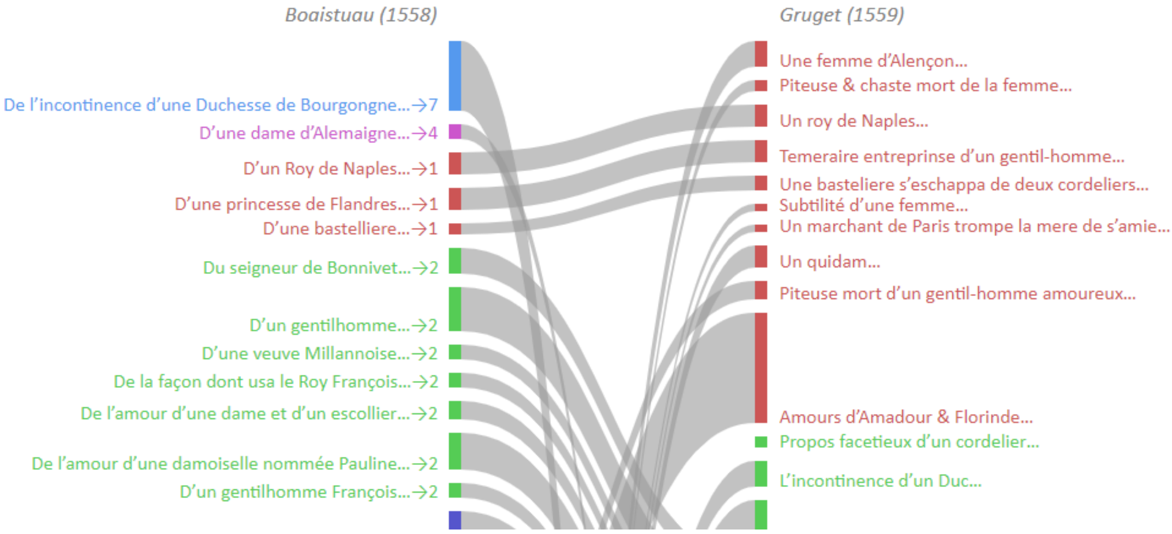


Figure 3.3: Sankey diagram built by SankeyCompare to compare two editions of the *Heptaméron* by Marguerite de Navarre, published in 1558 by Boaistuau and in 1559 by Gruget. Short stories are colored depending on which day of the 1559 edition they are associated with.

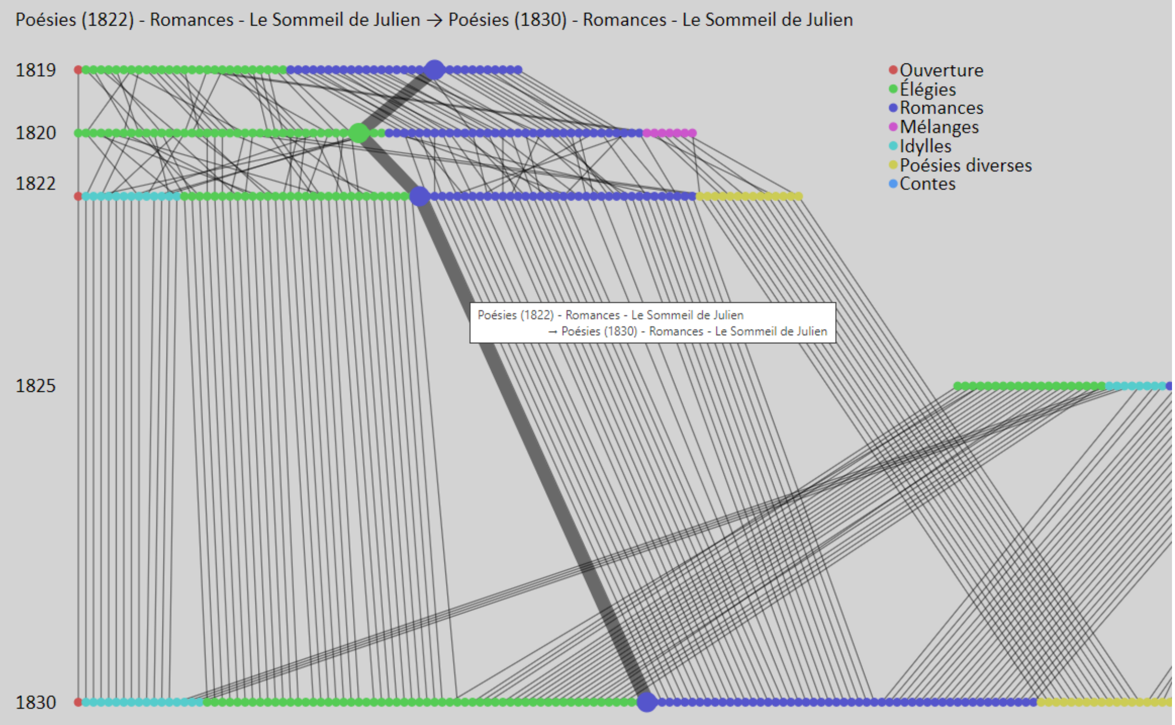


Figure 3.4: Overview of an interactive visualization of the first poetry books by Marceline Desbordes-Valmore published in 1819, 1820, 1822, 1825 and 1830. Each colored dot represents a poem. Two linked dots correspond to consecutive editions of a poem, details appear in a tooltipe when hovering over the link and clicking on the link redirects to an automatic alignment between the two versions of the poem.

3.2 Modernizing texts using text alignments

3.2.1 Learning modernization rules from a parallel corpus

In 2020, with Eleni Kogkitsidou, who joined the *Cité des dames*[7] project for one year as a postdoctoral researcher, we supervised the internship of Jonathan Poinhos, a first-year master student at Université Gustave Eiffel, to create a tool to automatically modernize old French texts. We actually targeted texts from the 17th century and after, and we designed an approach intensively based on text alignment. The *ABA method* (for Alignment-Based Approach), implemented in Python and available at https://github.com/johnseazer/aba, has two goals:

- to analyze a parallel corpus made of sentences in 17th century French and their normalized version in 21st century French, in order to identify frequently transformed sequences of characters;
- to deduce from this analysis a set of character transformation rules to automatically modernize a text from the 17th century or later.

[7] See section 2.4.2 for a presentation of this project.

First (GBG+22), the idea was to refine the alignment of the parallel corpus at sentence level into an alignment at word level, allowing to build a modernization dictionary (associating each word of the original version of the texts with the corresponding word in the modernized version). By refining it further into an alignment at character level, it was also possible to observe frequent transformations of sequences of consecutive characters. Those transformations were grouped into observation rules, such as the transformation of the suffixes “ois” and “oit” into “ais” and “ait”. These groups of observed transformation rules were also linked with published descriptions of the evolution of the French language (Pel95; Vac10) in (GGBS23).

Secondly, several frequent character transformation rules observed in the analysis described above (BPK+22) could actually be coded by Jonathan Poinhos as rewriting rules transforming portions of words (e. g. a final “ois” into “ais” to express the past tense of *imparfait*). So, for each word of the input text, after first checking if the word is present in the general large-coverage French lexicon Morphalou 3.1 (ATI23) (in which case we keep it), or if it is present in the replacement dictionary we learned from a parallel corpus of original and modernized versions of texts (then we replace it by the observed modernized version), we test every possible combinations of rewriting rules and for each obtained candidate result, we check in the Morphalou lexicon if it is present, in which case we output this candidate. If it does not apply to any of them, we keep the initial word. Although limited to words whose modernized version is present in the Morphalou lexicon, this technique nevertheless yields good results in practice, much better than word replacement method based on a collaborative word replacement dictionary used in Wikisource.

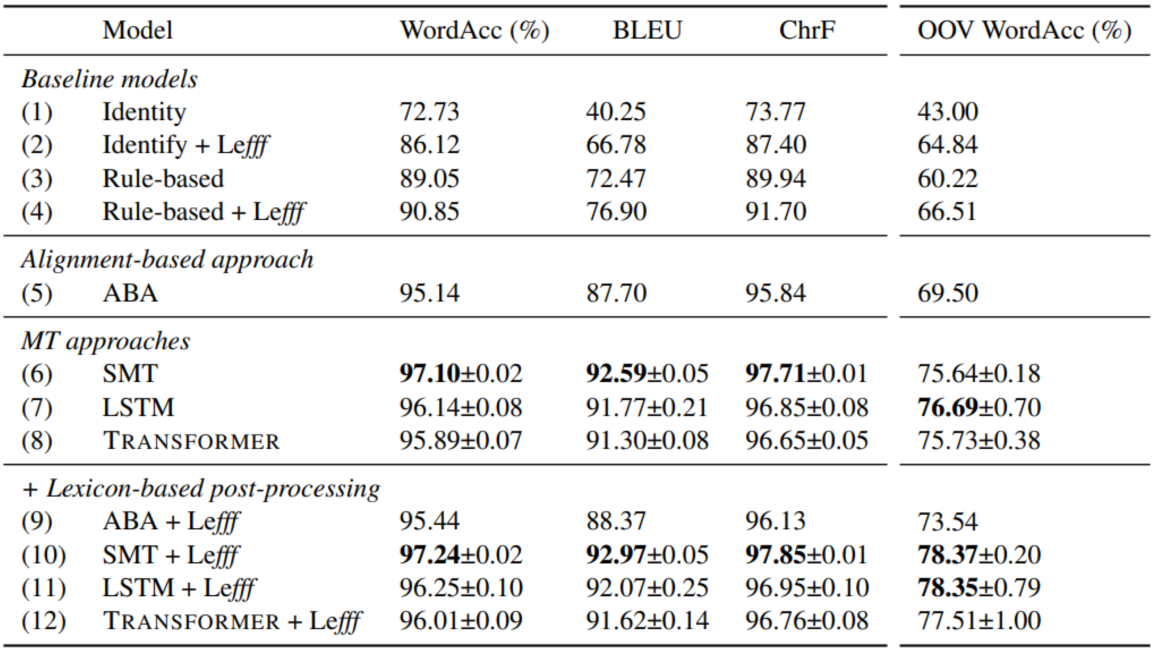
This approach was included in a study made with the coauthors who also worked with us on the description of the evolution of the French language, Simon Gabay, Rachel Bawden, Pedro Ortiz Suárez and Benoit Sagot (BPK+22). We obtained the results in Figure 3.5, which show that statistical machine translation obtains the best results for most evaluation criteria.

Being implemented as a simple Python script with few dependencies, tje ABA method was deployed on the web server of LIGM and is currently available at http://igm.univ-mlv.fr/~gambette/text-processing/aba.

3.2.2 Using modernization for named entity recognition

We also used the alignment-based modernization method ABA, presented in the previous section, for a study where we compare the performance of automatic methods for geographical named entity recognition on texts of the 17th century (KG20). Several methods are applied to the original and to the manually or automatically modernized versions of the texts. The obtained results are displayed in Figure 3.6. They show that the CasEN method developed in Tours, based on local grammar graphs implemented in Unitex, with a post-processing designed to avoid confusion between place names and person names, obtain the best results on manually modernized texts. This shows the importance of modernization in this case.

More recent results, which were obtained with a neural network model (BiLSTM-CRF)



	Model	WordAcc (%)	BLEU	ChrF	OOV WordAcc (%)
Baseline models					
(1)	Identity	72.73	40.25	73.77	43.00
(2)	Identify + *Lefff*	86.12	66.78	87.40	64.84
(3)	Rule-based	89.05	72.47	89.94	60.22
(4)	Rule-based + *Lefff*	90.85	76.90	91.70	66.51
Alignment-based approach					
(5)	ABA	95.14	87.70	95.84	69.50
MT approaches					
(6)	SMT	**97.10**±0.02	**92.59**±0.05	**97.71**±0.01	75.64±0.18
(7)	LSTM	96.14±0.08	91.77±0.21	96.85±0.08	**76.69**±0.70
(8)	Transformer	95.89±0.07	91.30±0.08	96.65±0.05	75.73±0.38
+ Lexicon-based post-processing					
(9)	ABA + *Lefff*	95.44	88.37	96.13	73.54
(10)	SMT + *Lefff*	**97.24**±0.02	**92.97**±0.05	**97.85**±0.01	**78.37**±0.20
(11)	LSTM + *Lefff*	96.25±0.10	92.07±0.25	96.95±0.10	**78.35**±0.79
(12)	Transformer + *Lefff*	96.01±0.09	91.62±0.14	96.76±0.08	77.51±1.00

Figure 3.5: Results of the modernization approaches on the test set. "+ Lefff" indicates that a lexicon-based post-processing was applied. Word accuracy is also evaluated on out-of-vocabulary tokens.

trained on 17th century texts, yield an F-measure of 0.84 for geographical named entities. This result, which is comparable to the one we obtained on another corpus of manually modernised 17th century texts suggests that it would be interesting to make an extended study of this entire range of methods for named entity recognition on early modern texts.

3.3 Other similarity models: uncovering the sources of classical French theater

Sometimes, similarity between two texts goes beyond the similarity between their sequences of words. This is particularly true for theater plays in cases where the author claims to have been inspired by another play, or chooses to adapt a play describing another context, or written in another language. This section presents new approaches taking into account structure similarity to detect such influences.

3.3.1 *Hyperpièces*, a corpus of classical plays and their sources

During her doctorate at Sorbonne Université, Céline Fournial had studied the sources of the French classical theater. More precisely, the appendix of her doctoral thesis (Fou19) listed more than 500 tragedies, comedies or tragicomedies, as well as the texts she had identified as sources. In 2020, I structured this data into a database and we designed a website to explore

	precision			recall			F-measure		
tool	original	auto	modern	original	auto	modern	original	auto	modern
CoreNLP	0,232	0,271	0,267	**0,799**	**0,814**	0,792	0,360	0,407	0,400
spaCy	0,363	0,412	0,443	0,701	0,731	0,754	0,479	0,527	0,558
SEM	0,485	0,652	0,649	0,610	0,602	0,636	0,540	0,626	0,642
PERDIDO	0,590	0,629	0,644	0,523	0,534	0,610	0,554	0,578	0,626
CasEN	0,743	0,723	0,742	0,602	0,602	**0,830**	0,665	0,657	0,784
CasEN+R	**0,843**	**0,814**	**0,879**	0,610	0,614	0,826	**0,708**	**0,700**	**0,852**

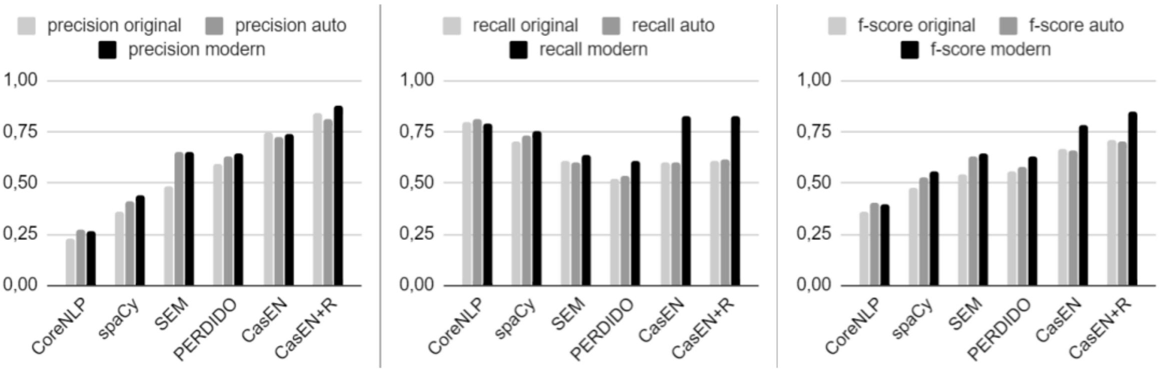


Figure 3.6: Precision, recall and F-measure obtained by 6 tools for geographical named entity recognition, respectively on the original corpus (“original”), the automatically modernised corpus (“auto”) and the manually modernised corpus (“modern”).

this corpus, *Hyperpièces*. I coded it as a free software in HTML, CSS and Javascript using Google Sheets as data sources, available at https://celinefournial.github.io/hyperpieces/. It uses Sankey diagram visualizations, shown in Figure 3.7, to illustrate the influences of specific categories of texts, namely French, Italian or Spanish plays, as well as biblic or antique texts, which inspired those plays. Focusing on the influences between French plays, I also coded the interactive visualization of the directed acyclic graph displaying those influences, where plays are displayed according to their publication date, illustrated in figure 3.8. Images of title pages of published plays were also included in the website, as well as editions of the plays, either in image or text version. An attempt to automatically look for scanned versions of the plays in Gallica only allowed to find a few dozen plays available online, due to variants between titles in the *Catalogue général de la Bibliothèque nationale de France* and in the *Hyperpièces* database. Similarly, automatically looking for plays encoded in XML-TEI available at *theatre-classique.fr* only resulted in a few identifications and had to be completed manually.

Two natural questions emerged from these visualisations:

- would it be possible to automatically detect influences between plays, based on some similiarities between them?
- would it be possible to characterize more precisely which kind of influence is observed between two plays?

A partial answer to these questions is provided on the website by giving the possibility to compare the character networks of two plays, and therefore to observe possible similarities between character names or interactions. In the *character networks* representation

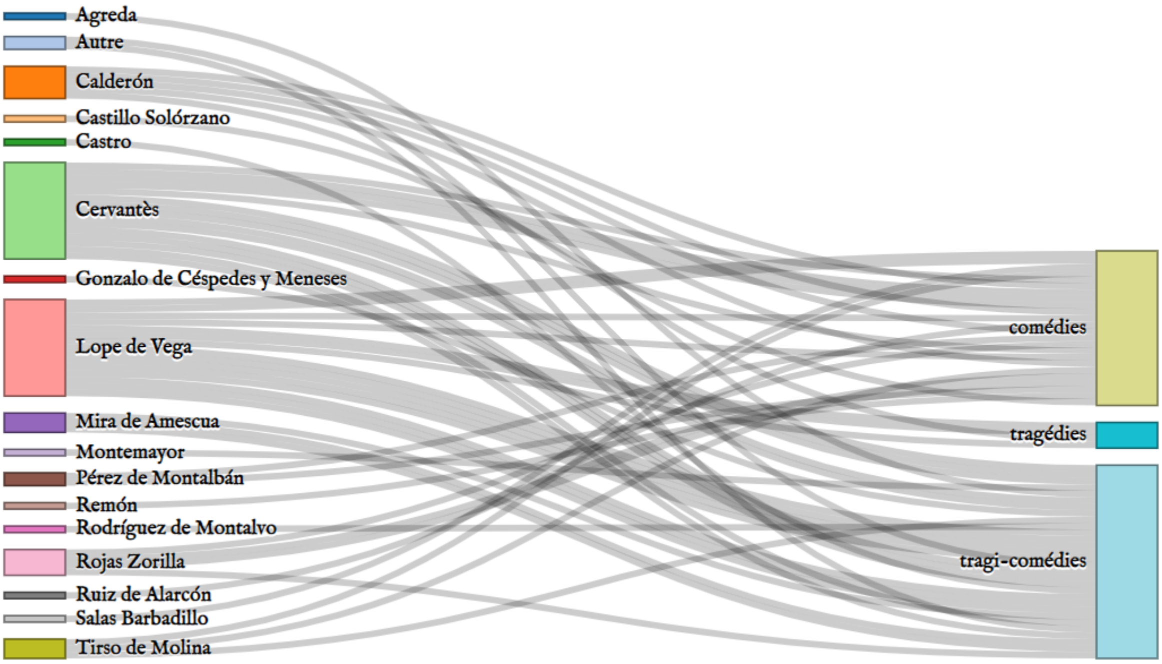


Figure 3.7: Sankey diagram displaying French plays from 1550 to 1650 with Spanish sources. The interactive version at https://celinefournial.github.io/hyperpieces/ also displays the name of both plays when the mouse hovers a link between them.

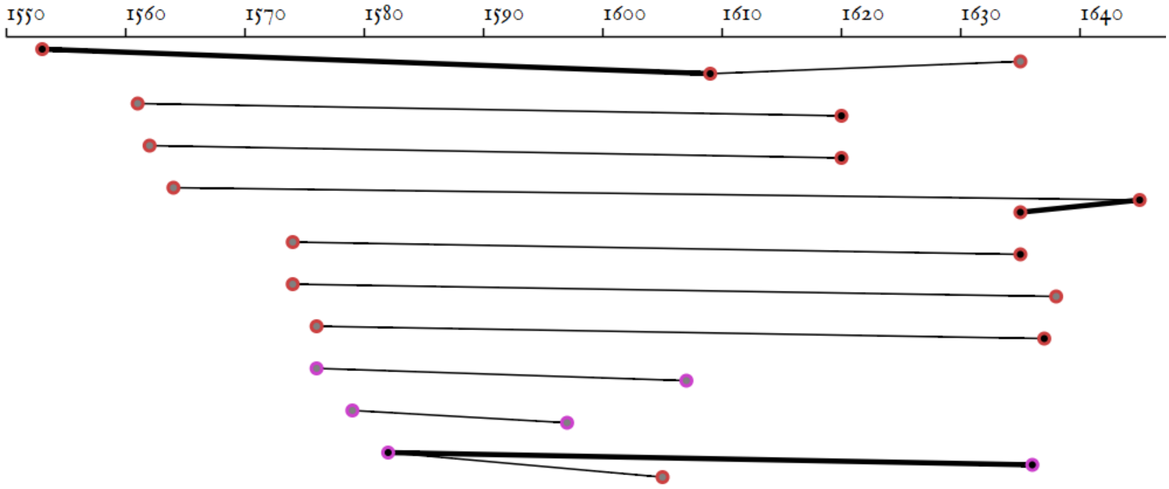


Figure 3.8: "Lineages" of French plays. The interactive version at https://celinefournial.github.io/hyperpieces/ also displays the name of both plays when the mouse hovers a link between them and compares the character networks on the two plays when clicking on the link between them.

of a play, the vertices of the network represent characters of the play and an edge (a, b) is present if characters a and b speak in the same scene. A *weighted directed character network* can be defined similarly, except arc (a, b) is weighted by the length of the text said by character a in scenes where character b also speaks. Those definitions allow to build such graphs, in practice, directly from the XML-TEI encoding (Bur14) of the theater plays, where each portion of the text is associated with the character who says it and with the scene where it is pronounced.

The *Hyperpièces* website includes character network visualizations coded by Frédéric Glorieux for the Dramagraphe project[8], which I modified to help compare plays encoded in XML-TEI by visually comparing their character networks. First, the vertices representing characters are displayed in a circle, in the order they appear in the play. Furthermore, arc (a, b) is colored red if character b is *dominated* by character a in the play, which was defined by Solomon Marcus as the fact that a appears in every scene where b also appears (we slightly adapted the definition to be able to automatically evaluate it with the information of our encoded texts, considering that b is dominated by a if a *speaks* in every scene where b *speaks*). These two modifications aim at highlighting other possible similarities between characters among two theater plays beyond their speaking relationships on scene, namely the fact that they appear in similar order, and the fact that the domination relationships are preserved.

3.3.2 Detecting and analyzing structure similarities between theater plays

As different possible ways of modeling theatre plays may be useful to detect different kinds of similarities, and possibly influences, I wrote a research internship offer, in Autumn 2020, to explore various kinds of models, based on sets of words, trees or networks, and design similarity detection approaches for those models. Aaron Boussidan, a student at ENS Paris Saclay, who had used the “césure” arrangement to interrupt his studies for one year to study theatre and practice acting, chose this internship. He also obtained funding from the ENS Paris-Saclay to start a doctoral project on models and methods for the detection of similarities between theater plays, with Cyril Nicaud as the doctoral director, myself as the main supervisor and Pierre Bourhis as a co-supervisor. Aaron worked both on a typology of similarities between them (sequel, prequel, full rewrite of the story, adaptation to another context, etc.) and on possible models to express various kinds of similarities between plays.

I had already been working, during my PhD, on a “distant reading” approach, to study the similarities between two theater plays by Corneille, *Cinna* and *Othon*, with Delphine Amstutz. We had used TreeCloud visualizations, as well as several functionalities of Lexico 3, to highlight differences between the two plays, which both deal with the links between love and power, but with different views. This methodology based on statistical representations of the texts, called “textométrie” in French, requires human interpretation of those representations to analyze the texts, I therefore wanted to develop more automated approaches.

In particular, the character matrices studied by Marc Douguet in his PhD thesis, which is similar to an earlier model introduced by Solomon Marcus (Pos74; BN74), seemed relevant, with possibly interesting algorithmic challenges for problems aiming at reordering the lines

[8] https://obvil.huma-num.fr/dramagraph/

of the matrix optimally to find similarities between them. Indeed, those binary matrices represent the presence or absence of characters in scenes of a play, the scenes being the columns of the matrix and the characters being its lines. This model therefore introduces a natural order among the characters (the order of the line), whereas it would be more relevant to consider unordered sets of characters per scene. Then, comparing two plays would consist, according to this model, in comparing two sequences of sets of characters, and therefore trying to find a mapping between characters such that some edit distance between the two sequences of sets of characters on the same alphabet is minimized. Looking for a simpler version of the problems, where sets are simply singletons, we realized that adopting this model would actually correspond to studying the succession of characters in the play. Maxime Crochemore told us that this problem of mapping characters of the alphabet of the first string to characters of the alphabet of the second string in such a way that the optimal edit distance between the two strings could be found had already been studied in the stringology literature, under the name "parameterized matching" (Bak99; HLS07; KKL09).

Let us give a more formal definition of the main parameterized matching problem we studied, PM^d and show the link with our application of interest.

Definition 1. *We consider an edit distance d, that is a function which, given two words, counts the minimum number of operations (chosen among insertion, i. e. adding a character in the string, deletion, i. e. removing a character in the string and substitution, i. e. replacing a character in the string) to obtain the second word starting from the first word.*

The Parameterized Matching problem under d, denoted by PM^d, is defined as:

- ***Input:** an integer k and two words u and v on the alphabet $\Pi \cup \Sigma$, where Π is called the* alphabet of parameters *and Σ is the* alphabet of constants.
- ***Problem:** Does there exist another word u' on the alphabet $\Pi \cup \Sigma$ such that $d(u, u') \leq k$ and u' and v are* parameterized matching, *i.e. there exists a one-to-one* character mapping function *$f : \Pi \cup \Sigma \to \Pi \cup \Sigma$ such that:*
  - $\forall c \in \Sigma, f(c) = c$,
  - $\forall c \in \Pi, f(c) \in \Pi$,
  - *and $f(u') = v$?*

An example of positive instance to $PM^{d_{DIS}}$, where deletions, insertions and substitutions are allowed as edit operations, in given in figure 3.9.

In (BBG23b), we study the computation complexity of several variants of this problem, depending on the operations allowed in the edit distance (see table 3.1). We obtain similar results for other variants of the problem, namely function matching, where the one-to-one requirement for the function in the definition of parameterized matching is dropped, which actually results in two different natural ways of defining the problem. To handle these NP-complete problems in practice, a MaxSAT encoding of the problem is provided

$$u = \text{TMTMTMTMTMMAUAUAUAUOUOJUJUJMJMJMJMJ}$$
$$v = \text{NMNMUMUMUMUMUMKUKUOUOKUUMUMUJUJUJJUJMJMJMJM}$$

(a)

```
TMTMTMTMTM_MAU_AUAUAUO______U_OJU_JUJMJMJMJMJ
====|=|=|=+=-=+=====|==+++++++=+|==+===========-
NMNMUMUMUMUM_UMKUKUOUOKUUMUMUJUJUJJUJMJMJMJM_
```

(b)

Figure 3.9: (a) The words u and v associated with the fifth act of the theater and opera versions of *Médée* by Pierre and Thomas Corneille respectively; the letter-character correspondence is the following: A: Créon, J: Jason, K: le chœur, M: Médée, N: Nérine, O: Cléone, T: Théodar, U: Créuse. (b) An alignment (deletions are denoted by -, insertions by + and substitutions by |) showing that considering alphabets $\Sigma = \{\text{J}, \text{M}, \text{O}, \text{U}\}$ and $\Pi = \{\text{A}, \text{K}, \text{N}, \text{T}\}$, the integer $k = 17$ and words u and v provide a positive instance of the $PM^{d_{DIS}}$ problem, with character mapping function f such that $f(\text{T}) = \text{N}$ and $f(\text{A}) = \text{K}$ and word $u' = \text{NMNMNMNMNMMKUKUKUKUOUOJUJUJMJMJMJMJ}$.

for the $PM^{d_{ID}}$ problem. It was coded by Aaron Boussidan using the MaxHS solver (Dav14) available at http://www.maxhs.org and I also coded a simple brute-force algorithm trying all possible bijections for the alphabet of parameters, which is therefore an FPT algorithm in the size of the alphabet of paremeters[9].

d	$\emptyset$	S
$\emptyset$	P (Bak99)	P (HLS07)
D	**NP-complete (Th. 12 of (BBG23b))**	**NP-complete (Corr. 14 of (BBG23b))**
I	**NP-complete (Cor. 14 of (BBG23b))**	**NP-complete (Cor. 14 of (BBG23b))**
DI	NP-complete (KKL09)	**NP-complete (Th. 13 of (BBG23b))**

Table 3.1: Complexity of the variants of parameterized matching PM^d, depending on the kind of operations (D: deletion, I: insertion, S: substitution) allowed in the edit distance d.

This parameterized matching approach seems relevant for adaptations of plays from one language to another, as character names are sometimes changed in the process. In this case, the optimal alignment found outputs a character mapping function which may identify the right characters between the two plays, but an accurate alignment of the lines spoken by the characters requires to take into account the semantic content of these lines.

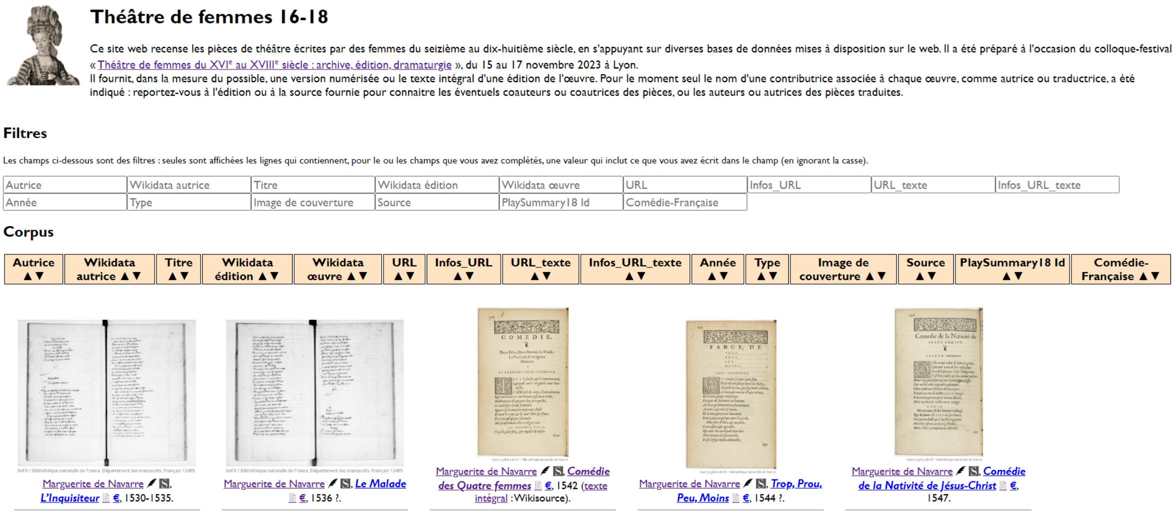
In (BBFG24), we focus on this after identifying a few pairs of plays whose parameterized distance is quite low.

[9] https://github.com/AaronFive/paramatch/

3.3.3 Towards future work on early modern plays written by women

I was invited in 2023 to provide an overview on the availability of digital versions of early modern plays written by women, by the organizing team of the workshop *Théâtre de femmes du XVI^e au XVIII^e siècle : archive, édition, dramaturgie*. Using my previous experiences with the *Hyperpièces* project (see section 3.3.1) and with the #Theatrices2022 project of association Le deuxième texte for which I had collected more than 200 theater plays written by women[10], I started to develop a database of early modern plays written by women, using other databases as sources such as *Le théâtre français de la Révolution à l'Empire 1789-1815*[11], the *Répertoire du théâtre français imprimé au XVII^e siècle*[12], the *Table des femmes auteurs, entrepreneurs et salonnières* by David Trott[13] or the *Projet Play Summary 18*[14].

The website displaying this database is illustrated in figure 3.10. The references of more than 380 plays written by a total of more than 100 women writers, which can be associated to more that 180 scanned versions digital books, with more than 60 whose text is also available online. Each play was also associated with a Wikidata item, in order to encourage reusability of this data and future studies on this corpus of plays written by women.



Théâtre de femmes 16-18

Ce site web recense les pièces de théâtre écrites par des femmes du seizième au dix-huitième siècle, en s'appuyant sur diverses bases de données mises à disposition sur le web. Il a été préparé à l'occasion du colloque-festival « Théâtre de femmes du XVI^e au XVIII^e siècle : archive, édition, dramaturgie », du 15 au 17 novembre 2023 à Lyon.
Il fournit, dans la mesure du possible, une version numérisée ou le texte intégral d'une édition de l'œuvre. Pour le moment seul le nom d'une contributrice associée à chaque œuvre, comme autrice ou traductrice, a été indiqué : reportez-vous à l'édition ou à la source fournie pour connaitre les éventuels coauteurs ou coautrices des pièces, ou les auteurs ou autrices des pièces traduites.

Filtres

Les champs ci-dessous sont des filtres : seules sont affichées les lignes qui contiennent, pour le ou les champs que vous avez complétés, une valeur qui inclut ce que vous avez écrit dans le champ (en ignorant la casse).

Autrice	Wikidata autrice	Titre	Wikidata édition	Wikidata œuvre	URL	Infos_URL	URL_texte	Infos_URL_texte
Année	Type	Image de couverture	Source	PlaySummary18 Id	Comédie-Française			

Corpus

Autrice ▲▼	Wikidata autrice ▲▼	Titre ▲▼	Wikidata édition ▲▼	Wikidata œuvre ▲▼	URL ▲▼	Infos_URL ▲▼	URL_texte ▲▼	Infos_URL_texte ▲▼	Année ▲▼	Type ▲▼	Image de couverture ▲▼	Source ▲▼	PlaySummary18 Id ▲▼	Comédie-Française ▲▼

Marguerite de Navarre ✒ S, *L'Inquisiteur* 📄 €, 1530-1535.

Marguerite de Navarre ✒ S, *Le Malade* 📄 €, 1536 ?.

Marguerite de Navarre ✒ S, *Comédie des Quatre femmes* 📄 €, 1542 (texte intégral :Wikisource).

Marguerite de Navarre ✒ S, *Trop, Prou, Peu, Moins* 📄 €, 1544 ?.

Marguerite de Navarre ✒ S, *Comédie de la Nativité de Jésus-Christ* 📄 €, 1547.

Figure 3.10: An overview of the *Théâtre de femmes 16-18* website, available at https://citedesdames.github.io/theatre1618.

[10] https://george2etexte.wordpress.com/2022/03/11/les-femmes-savantes-precieuses-a-plus-dun-titre/
[11] https://theatre1789-1815.e-monsite.com/
[12] https://repertoiretheatreimprime.othone.org/
[13] http://homes.chass.utoronto.ca/~trott/fem_aut.htm
[14] https://web.archive.org/web/20220513122818/http://playsummary18.org/

Conclusion and future work

L'avenir m'a promis de riantes pensées.

Marceline Desbordes-Valmore, « Le retour à Bordeaux »,
Élégies et poésies nouvelles, Paris, Ladvocat, 1825.

This document summarizes my main research results on topics related to similarity, proximity and heredity, in bioinformatics, mostly in the domain of phylogenetic networks, and in digital humanities, where I show a special interest in alignment algorithms. For digital humanities, it illustrates how I have tried to get a full grasp of this research domain I had partially started to explore during my doctorate.

I have collaborated with several researchers in the humanities to fully understand their needs, not always being able to answer them with digital solutions, but sometimes being able to develop a useful tool or to extract an interesting theoretical problem for a theoretical study. I think that instead of simply trying to apply methods developed for biological sequences to texts, it is more relevant to get a good undestanding of the specificities of the input data in order to identify the right approach. This may lead to use the toolbox of algorithmics for bioinformatics or natural language processing, or to automatize classical approaches developed in the humanities, or uncover convergences between the two.

Instead of performing hands-on bench work related with the acquisition of biological sequences, I have developed several skills in acquiring, handling and making available textual data for projects in literature. In particular, I'm happy that my efforts to gather data, in the last five years, on the works of Marceline Desbordes-Valmore, helped to uncover texts which were unknown to specialists, and to gather a corpus which will open new research perspectives on her works, for example for genetic editing, or about the links between her poetry and music. Making tools available for the team currently working on a paper edition of her poetry works, supervised by Christine Planté and Andrea Schellino, is also a good way of both getting to understand the needs of the team and observing if the implemented functionalities are fully adequate solutions.

Working on the cultural heritage of women has convinced me that there is a lot to explore in many understudied texts written by women in the public domain. It is particularly exciting to be able to develop new digital tools and to test them on corpora where it may be easier to quickly find interesting and new results.

Being involved in open science, in particular through my mission at Université Paris-Est Marne-la-Vallée in 2019 and then at Université Gustave Eiffel in 2022 and 2023, I am also making some efforts to make my research data available under open licenses. I also contribute to wikis related with the Wikimedia Foundation, for example Wikidata, Wikimedia Commons or Wikisource, which I also use in my research projects. This approach also opens new project ideas for the years to come. For example, the automatic evaluation of the quality of proofread texts on Wikisource is an interesting question with several aspects to study. It could benefit from a detailed analysis of variants spotted by automatic text alignment algorithms between successive versions of proofread pages to be able to characterize the corrected errors. But other issues such as the choice of edition to make available may also be studied using data anlysis. Furthermore, a more extensive use of data from Wikidata for text mining or text analysis, using recently developed representations, such as graph embeddings, or clever ways of using genealogical data (for historical texts in particular), also seems promising.

Several years of involvment in CJC, Confédération des Jeunes Chercheurs and in ANDès, Association nationale des docteurs, have convinced me that supervising research, and most of all, people who do research, is not an easy task. I knew, in theory, that trying to reproduce the ideal conditions offered to me by people who supervised my research activity, Denis Bertrand, Olivier Gascuel, Daniel Huson, Michel Habib, Vincent Berry, Christophe Paul and Alain Guénoche, may not be sufficient with everyone. But I was able to experiment it in practice, and finding solutions when problems occur is not an easy task, but I'm probably able to anticipate problems now, more so than 12 years ago. It is even harder when time is limited by the amount of emergencies to deal with every day, but I hope I will continue improving my organization and time management skills.

In addition to these issues, encouraging diversity in the team of people I am interacting with is also important for me. So far, I have mostly been able to experiment it when hiring interns, which was more frequent than hiring PhD candidates or postdocs. Taking risks, sometimes, for example with foreign students or with a few students who had some difficulties in class, has so far mostly been rewarded with very nice results and sometimes impressive realizations. The challenge for the years to come will therefore be to continue attracting brilliant early-stage researchers, like Chuanming Dong, Eleni Kogkitsidou and Aaron Boussidan, to work on theoretical or applied topics in digital humanities. This will be essential to continue to develop new approaches, based on powerful algorithmic ideas, in order to obtain practical and easy-to-use tools and methods, taking advantage of the ones already developed over the last ten years.

Bibliography

[AG10] Delphine Amstutz and Philippe Gambette. Utilisation de la visualisation en nuage arboré pour l'analyse littéraire. In *JADT'10: 10th International Conference on statistical analysis of textual data*, page 12, Rome, Italy, June 2010. https://hal-lirmm.ccsd.cnrs.fr/lirmm-00448436. 28

[AGM16] Tushar Agarwal, Philippe Gambette, and David Morrison. Who is who in phylogenetic networks: Articles, authors and programs. Working paper, https://hal-upec-upem.archives-ouvertes.fr/hal-01376483, 2016. 7

[ASSU81] Alfred V. Aho, Yehoshua Sagiv, Thomas G. Szymanski, and Jeffrey D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing*, 10(3):405–421, 1981. http://doi.org/10.1137/0210030. 19

[ATI23] ATILF. Morphalou, 2023. ORTOLANG (Open Resources and TOols for LANGuage) – https://www.ortolang.fr/market/lexicons/morphalou. 51

[Bak99] Brenda S. Baker. Parameterized diff. In *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '99, page 854–855, USA, 1999. Society for Industrial and Applied Mathematics. https://dl.acm.org/doi/pdf/10.5555/314500.314968. 56, 57

[BBFG24] Pierre Bourhis, Aaron Boussidan, Céline Fournial, and Philippe Gambette. Detecting semantic or structural similarities for theater play comparison. In *JADT 2024, Mots comptés, textes déchiffrés*, Proceedings of the 17th International Conference on the Statistical Analysis of Textual Data, pages 139–148, Brussels, Belgium, June 2024. https://univ-eiffel.hal.science/hal-04722464. 57

[BBG+23a] Sarah J. Berkemer, Pierre Bourhis, Philippe Gambette, Lionel Seinturier, and Marion Tommasi. A database approach to solve the tree containment problem in phylogenetic networks. In *Mathematics of Evolution-Phylogenetic Trees and Networks, Workshop 1: Foundations of Networks*, 2023. 17

[BBG23b] Pierre Bourhis, Aaron Boussidan, and Philippe Gambette. On distances between words with parameters. In Laurent Bulteau and Zsuzsanna Lipták, editors, *CPM 2023*, volume 259 of *Proceedings of the 34th Annual Symposium on Combinatorial Pattern Matching*, pages 6:1–6:23, Champs-sur-Marne, Marne-la-Vallée, France, June 2023. Schloss Dagstuhl. http://doi.org/10.4230/LIPIcs.CPM.2023.6. 56, 57

[BCP23] Jean Barré, Jean-Baptiste Camps, and Thierry Poibeau. Operationalizing canonicity: A quantitative study of french 19th and 20th century literature. *Journal of Cultural Analytics*, 8(3), 2023. https://doi.org/10.22148/001c.88113. 41

[BDG+18] Christine Barats, Anne Dister, Philippe Gambette, Jean-Marc Leblanc, and Marie Pérès. Appeler à signer une pétition en ligne : caractéristiques linguistiques des appels. In *JADT 2018*, Proceedings of the 14th International Conference on the Statistical

Analysis of Textual Data, pages 68–75, Rome, Italy, June 2018. https://hal-upec-upem.archives-ouvertes.fr/hal-01775267. 30

[BGLL08] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. https://doi.org/10.1088/1742-5468/2008/10/P10008. 27

[BGS22] Laurent Bulteau, Philippe Gambette, and Olga Seminck. Reordering a tree according to an order on its leaves. In *CPM 2022*, volume 223 of *LIPIcs*, pages 24:1–24:15, Prague, Czech Republic, June 2022. Schloss Dagstuhl - Leibniz-Zentrum für Informatik. http://doi.org/http://doi.org/10.4230/LIPIcs.CPM.2022.24. 38

[BL87] Jean-Pierre Barthélemy and Nhuan Xuan Luong. Sur la topologie d'un arbre phylogénétique : aspects théoriques, algorithmes et applications à l'analyse de données textuelles. *Mathématiques et sciences humaines*, 100:57–80, 1987. http://www.numdam.org/item/MSH_1987__100__57_0/. 33

[Bla27] Sandrine Blanchard. Nos petites madeleines. *Le Monde*, 2013-03-27. https://www.lemonde.fr/idees/article/2013/03/27/nos-petites-madeleines_3148785_3232.html. 19

[BLS99] Andreas Brandstädt, Van Bang Le, and Jeremy P. Spinrad. *Graph classes: a survey*. SIAM, 1999. https://epubs.siam.org/doi/book/10.1137/1.9780898719796. 9

[BN74] Barron Brainerd and Victoria Neufeldt. On Marcus' methods for the analysis of the strategy of a play. *Poetics*, 3(2):31–74, 1974. https://doi.org/10.1016/0304-422X(74)90013-8. 55

[Bod86] Thierry Bodin. Marceline Desbordes-Valmore et ses musiciens. In *Actes du colloque Marceline Desbordes-Valmore et son temps 26 avril 1986*. Mémoires de la Société d'Agriculture, Sciences et Arts de Douai, 1986. 47

[BPK$^+$22] Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. Automatic normalisation of early modern french. In *LREC 2022 - 13th Language Resources and Evaluation Conference*, Marseille, France, June 2022. European Language Resources Association. https://hal.inria.fr/hal-03540226. 51

[BS16] Magnus Bordewich and Charles Semple. Reticulation-visible networks. *Advances in Applied Mathematics*, 78:114–141, 2016. https://doi.org/10.1016/j.aam.2016.04.004. 23

[BSS06] Mihaela Baroni, Charles Semple, and Mike Steel. Hybrids in real time. *Systematic biology*, 55(1):46–56, 2006. https://pubmed.ncbi.nlm.nih.gov/16507523/. 18

[Bun74] Peter Buneman. A note on the metric properties of trees. *Journal of Combinatorial Theory Series B*, 17(1):48–50, 1974. https://doi.org/10.1016/0095-8956(74)90047-1. 30

[Bur14] Lou Burnard. *What is the Text Encoding Initiative?* OpenEdition Press, 2014. https://books.openedition.org/oep/426. 55

[CBB^{+}14] Rosa Cetro, Marc M. Barbier, Philippe P. Breucker, Hilde Eggermont, Philippe Gambette, Tita Kyriacopoulou, Xavier Le Roux, Claude Martineau, and Nicolas N. Turenne. Vers une approche semi-automatique pour la définition de motifs d'argumentation utilisés dans les résumés de projets scientifiques du domaine de la biodiversité. *Revue des Nouvelles Technologies de l'Information*, RNTI-SHS-2(2):47–80, 2014. https://hal-upec-upem.archives-ouvertes.fr/hal-01090607. 29

[CG09] Christophe Crespelle and Philippe Gambette. Efficient neighbourhood encoding for interval graphs and permutation graphs and $O(n)$ breadth-first search. In *IWOCA'09: 20th International Workshop on Combinatorial Algorithms*, volume 5874 of *Lecture Notes in Computer Science*, pages 146–157, Hradec nad Moravicí, Czech Republic, June 2009. Springer Berlin / Heidelberg. https://hal-lirmm.ccsd.cnrs.fr/lirmm-00415935. 26

[CG13] Christophe Crespelle and Philippe Gambette. Linear-time constant-ratio approximation algorithm and tight bounds for the contiguity of cographs. In Ghosh, Subir Kumar, Tokuyama, and Takeshi, editors, *Seventh International Workshop on Algorithms and Computation*, volume 7748 of *Lecture Notes in Computer Science*, pages 126–136, Kharagpur, India, February 2013. Springer. https://hal.inria.fr/hal-00755257. 26

[CG14] Christophe Crespelle and Philippe Gambette. (nearly-)tight bounds on the contiguity and linearity of cographs. *Theoretical Computer Science*, 522:1–12, 2014. https://hal-upec-upem.archives-ouvertes.fr/hal-00915069. 26

[Che05] Sergiu Chelcea. BibAdmin 0.5, 2005. Software available at https://web.archive.org/web/20071103133222/https://gforge.inria.fr/projects/bibadmin/. 8

[CLPP16] Christophe Crespelle, Tien-Nam Le, Kevin Perrot, and Thi Ha Duong Phan. Linearity is strictly more powerful than contiguity for encoding graphs. *Discrete Mathematics*, 339(8):2168–2177, 2016. https://hal.science/hal-01424428. 26

[Con01] Jean-Gabriel Contamin. *Contribution à une sociologie des usages pluriels des formes de mobilisation : l'exemple de la pétition en France*. PhD thesis, Université Paris 1, 2001. http://www.theses.fr/2001PA010329. 30

[Cot16] Jérôme Cottanceau. *Le choix du meilleur urinoir ; et 19 autres délicats problèmes qui prouvent que les maths servent à quelque chose !* Belin, 2016. 19

[CRV08] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. A perl package and an alignment tool for phylogenetic networks. *BMC Bioinformatics*, 9:1–5, 2008. https://doi.org/10.1186/1471-2105-9-175. 12

[CRV09] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. Comparison of tree-child phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):552–569, 2009. https://doi.org/10.1109/TCBB.2007.70270. 15

[Dav14] Jessica Davies. *Solving MAXSAT by Decoupling Optimization and Satisfaction*. PhD thesis, University of Toronto, Canada, 2014. http://hdl.handle.net/1807/43539. 57

[DGD21] Chuanming Dong, Philippe Gambette, and Catherine Dominguès. Extracting event-related information from a corpus regarding soil industrial pollution. In *KDIR*

2021, volume 1 of *13th International Conference on Knowledge Discovery and Information Retrieval*, pages 217–224, Setúbal, Portugal, October 2021. SciTePress. https://hal.archives-ouvertes.fr/hal-03366097. 41

[DGD22] Chuanming Dong, Philippe Gambette, and Catherine Dominguès. Extraction et caractérisation de noyaux d'événements liés à la pollution industrielle. In *JADT 2022*, pages 354–360, 2022. https://hal.science/hal-04225005v1. 41

[DHK+11] Andreas Dress, Katharina Huber, Jacobus Koolen, Vincent Moulton, and Andreas Spillner. *Basic Phylogenetic Combinatorics*. Cambridge University Press, 2011. 21

[Don23] Chuanming Dong. *Construction d'une mémoire des sites potentiellement pollués à l'aide de méthodes de traitement automatique des langues*. PhD thesis, Université Gustave Eiffel, 2023. https://hal.science/tel-04500937. 40

[dR+] Hendrik Nicolaas de Ridder et al. Information System on Graph Classes and their Inclusions (ISGCI). https://www.graphclasses.org. 9

[ERK16] Maciej Eder, Jan Rybicki, and Mike Kestemont. Stylometry with R: a package for computational text analysis. *The R Journal*, 8(1), 2016. https://doi.org/10.32614/RJ-2016-007. 35

[Eve16] Martin Paul Eve. "You have to keep track of your changes": The version variants and publishing history of David Mitchell's *Cloud Atlas*. *Open Library of Humanities*, 2(2):1–34, 2016. https://doi.org/10.16995/olh.82. 48

[Fau14] Charlotte Fauve. Les bâtisseurs de la « cité des dames ». *Télérama*, 3866:38, 2024-02-14. 42

[Fel03] Joseph Felsenstein. *Inferring phylogenies*. Sinauer Associates, 2003. 20

[FHW80] Steven Fortune, John Hopcroft, and James Wyllie. The directed subgraph homeomorphism problem. *Theoretical Computer Science*, 10(2):111–121, 1980. https://doi.org/10.1016/0304-3975(80)90009-2. 15, 16

[FKP15] Jittat Fakcharoenphol, Tanee Kumpijit, and Attakorn Putwattana. A faster algorithm for the tree containment problem for binary nearly stable phylogenetic networks. In *Proceedings of the The 12th International Joint Conference on Computer Science and Software Engineering (JCSSE'15)*, pages 337–342. IEEE, 2015. http://dx.doi.org/10.1109/JCSSE.2015.7219820. 15

[Fou19] Céline Fournial. *Imitation et création dans le "théâtre moderne" (1550-1650) : la question des cycles d'inspiration*. PhD thesis, Sorbonne Université, 2019. http://www.theses.fr/2019SORUL012. 52

[Gam09] Philippe Gambette. Mathématiques des papillotes (1/2), 2009. Blog *Je véronise*, https://gambette.blogspot.com/2009/11/mathematiques-des-papillotes.html. 19

[Gam10] Philippe Gambette. *Méthodes combinatoires de reconstruction de réseaux phylogénétiques*. Theses, Université Montpellier II - Sciences et Techniques du Languedoc, November 2010. https://tel.archives-ouvertes.fr/tel-00608342. 6

[Gam23] Philippe Gambette. Phylogenetic networks found in scientific publications, 2023. http://doi.org/10.57745/VIW7B2. 13

[GBG+22] Simon Gabay, Rachel Bawden, Philippe Gambette, Jonathan Poinhos, Eleni Kogkitsidou, and Benoît Sagot. Le changement linguistique au XVII^e^ siècle : nouvelles approches scriptométriques. In *CMLF 2022 - 8^e^ Congrès Mondial de Linguistique Française*, volume 138 of *SHS Web of conferences*, pages 02006.1–14, Orléans, France, July 2022. EDP Sciences. http://doi.org/10.1051/shsconf/202213802006. 51

[GBP12] Philippe Gambette, Vincent Berry, and Christophe Paul. Quartets and unrooted phylogenetic networks. *Journal of Bioinformatics and Computational Biology*, 10(4):1250004.1–1250004.23, 2012. https://hal-upec-upem.archives-ouvertes.fr/hal-00678046. 6, 14, 20

[GDZ17] Andreas Gunawan, Bhaskar DasGupta, and Louxin Zhang. A decomposition theorem and two algorithms for reticulation-visible networks. *Information and Computation*, 252:161–175, 2017. http://dx.doi.org/10.1016/j.ic.2016.11.001. 15

[GELR14] Philippe Gambette, Hilde Eggermont, and Xavier Le Roux. Temporal and geographical trends in the type of biodiversity research funded on a competitive basis in European countries, 2014. BiodivERsA report, https://www.biodiversa.org/700/download. 29

[GFL04] Jean-Gabriel Ganascia, Irène Fenoglio, and Jean-Louis Lebrave. EDITE MEDITE : un logiciel de comparaison de versions. *Actes de JADT*, pages 468–478, 2004. http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_044.pdf. 48

[GG02] Alain Guénoche and Henri Garetta. Representation and evaluation of partitions. In *Classification, clustering and data analysis, Proceedings of IFCS'02*, 2002. https://doi.org/10.1007/978-3-642-56181-8_14. 30, 31

[GG11] Philippe Gambette and Alain Guénoche. Bootstrap clustering for graph partitioning. *RAIRO - Operations Research*, 45(4):339–352, October 2011. https://hal-upec-upem.archives-ouvertes.fr/hal-00676989. 27

[GGBP+17] Annie Glatigny, Philippe Gambette, Alexa Bourand-Plantefol, Geneviève Dujardin, and Marie-Hélène Mucchielli-Giorgi. Development of an in silico method for the identification of subcomplexes involved in the biogenesis of multiprotein complexes in saccharomyces cerevisiae. *BMC Systems Biology*, 11(67):1–12, 2017. https://hal-upec-upem.archives-ouvertes.fr/hal-01560671. 27

[GGBS23] Simon Gabay, Philippe Gambette, Rachel Bawden, and Benoît Sagot. Ancien ou moderne ? pistes computationnelles pour l'analyse graphématique des textes écrits au XVII^e^ siècle. *Linx*, 85, February 2023. http://doi.org/10.4000/linx.9346. 51

[GGE+18] Lise Goudeseune, Philippe Gambette, Hilde Eggermont, André Heughebaert, and Xavier Le Roux. The BiodivERsA database: a mapping of research on biodiversity and ecosystem services in Europe over 2005-2015, 2018. BiodivERsA report, https://www.biodiversa.org/1655/download. 29

[GGKS95] Paul W. Goldberg, Martin C. Golumbic, Haim Kaplan, and Ron Shamir. Four strikes against physical mapping of DNA. *Journal of Computational Biology*, 2(1):139–152, 1995. https://doi.org/10.1089/cmb.1995.2.139. 26

[GGL+15a] Philippe Gambette, Andreas D.M. Gunawan, Anthony Labarre, Stéphane Vialette, and Louxin Zhang. Locating a tree in a phylogenetic network in quadratic time. In *RECOMB 2015*, volume 9029 of *LNCS*, pages 96–107, Varsovie, Poland, April 2015. Springer. https://hal-upec-upem.archives-ouvertes.fr/hal-01116231. 15, 23

[GGL+15b] Philippe Gambette, Andreas D.M. Gunawan, Anthony Labarre, Stéphane Vialette, and Louxin Zhang. Solving the tree containment problem for genetically stable networks in quadratic time. In Zsuzsanna Lipták and William F. Smyth, editors, *IWOCA 2015*, volume 9538 of *Proceedings of the 26th International Workshop on Combinatorial Algorithms*, pages 197–208, Verona, Italy, October 2015. Springer. https://hal-upec-upem.archives-ouvertes.fr/hal-01226035. 15

[GGL+18] Philippe Gambette, Andreas D.M. Gunawan, Anthony Labarre, Stéphane Vialette, and Louxin Zhang. Solving the tree containment problem in linear time for nearly stable phylogenetic networks. *Discrete Applied Mathematics*, 246:62–79, 2018. https://hal-upec-upem.archives-ouvertes.fr/hal-01575001. 15

[GGN12] Philippe Gambette, Núria Gala, and Alexis Nasr. Longueur de branches et arbres de mots. *Corpus*, 11(-):129–146, 2012. https://hal-upec-upem.archives-ouvertes.fr/hal-00822993. 30

[GH12] Philippe Gambette and Katharina Huber. On encodings of phylogenetic networks of bounded level. *Journal of Mathematical Biology*, 65(1):157–180, 2012. https://hal.archives-ouvertes.fr/hal-00609130. 21, 22

[GHK17] Philippe Gambette, Katharina Huber, and Steven Kelk. On the challenge of reconstructing level-1 phylogenetic networks from triplets and clusters. *Journal of Mathematical Biology*, 74(7):1729–1751, 2017. https://hal-upec-upem.archives-ouvertes.fr/hal-01391430. 21

[GHS17] Philippe Gambette, Katharina Huber, and Guillaume Scholz. Uprooted phylogenetic networks. *Bulletin of Mathematical Biology*, 79(9):2022–2048, 2017. https://hal-upec-upem.archives-ouvertes.fr/hal-01570943. 21

[GM13] Philippe Gambette and William Martinez. L'affaire du Mediator au prisme de la textométrie. *Texto ! Textes et Cultures*, XVIII(4):3318.1–3318.9, 2013. https://hal-upec-upem.archives-ouvertes.fr/hal-00881639. 28

[GSLP21] Philippe Gambette, Olga Seminck, Dominique Legallois, and Thierry Poibeau. Evaluating hierarchical clustering methods for corpora with chronological order. In *EADH2021: Interdisciplinary Perspectives on Data. Second International Conference of* *the European Association for Digital Humanities*, Krasnoyarsk, Russia, September 2021. EADH. https://hal.archives-ouvertes.fr/hal-03341803. 36

[Gué11] Alain Guénoche. Consensus of partitions: a constructive approach. *Advances in Data Analysis and Classification*, 5(3):215–229, October 2011. http://doi.org/10.1007/s11634-011-0087-6. 27

[Gun18] Andreas Gunawan. Solving the tree containment problem for reticulation-visible networks in linear time. In *5th International Conference on Algorithms for Computational Biology (AlCoB 2018)*, volume 10849 of *LNCS*, pages 24–36. Springer, 2018. https://doi.org/10.1007/978-3-319-91938-6_3. 15

[Gus97] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge University Press, 1997. 2

[GV09] Philippe Gambette and Jean Véronis. Visualising a text with a tree cloud. In *IFCS'09: International Federation of Classification Societies Conference*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 561–569, Dresde, Germany, March 2009. Springer Berlin / Heidelberg. https://hal-lirmm.ccsd.cnrs.fr/lirmm-00373643. 28

[GvIJ+17] Philippe Gambette, Leo van Iersel, Mark Jones, Manuel Lafond, Fabio Pardi, and Celine Scornavacca. Rearrangement moves on rooted phylogenetic networks. *PLoS Computational Biology*, 13(8):e1005611, 2017. http://doi.org/10.1371/journal.pcbi.1005611. 17, 20

[GZ15] Andreas DM Gunawan and Louxin Zhang. Bounding the size of a network defined by visibility property. *arXiv preprint*, 2015. https://arxiv.org/abs/1510.00115. 23

[HLMW16] Katharina T. Huber, Simone Linz, Vincent Moulton, and Taoyang Wu. Spaces of phylogenetic networks from generalized nearest-neighbor interchange operations. *Journal of Mathematical Biology*, 72(3):699–725, 2016. https://doi.org/10.1007/s00285-015-0899-7. 19, 20, 22

[HLS07] Carmit Hazay, Moshe Lewenstein, and Dina Sokol. Approximate parameterized matching. *ACM Trans. Algorithms*, 3(3):29–es, 2007. http://doi.org/10.1145/1273340.1273345. 56, 57

[HMSW16] Katharina T Huber, Vincent Moulton, Mike Steel, and Taoyang Wu. Folding and unfolding phylogenetic trees and networks. *Journal of Mathematical Biology*, 73:1761–1780, 2016. https://pubmed.ncbi.nlm.nih.gov/27107869/. 14

[HMW16] Katharina T. Huber, Vincent Moulton, and Taoyang Wu. Transforming phylogenetic networks: Moving beyond tree space. *Journal of Theoretical Biology*, 404:30–39, 2016. https://doi.org/10.1016/j.jtbi.2016.05.030. 19

[HRS11] Daniel H. Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications.* Cambridge University Press, 2011. 14, 15

[JNST06] Guohua Jin, Luay Nakhleh, Sagi Snir, and Tamir Tuller. Maximum likelihood of phylogenetic networks. *Bioinformatics*, 22(21):2604–2611, 2006. https://doi.org/10.1093/bioinformatics/btl452. 19

[KG20] Eleni Kogkitsidou and Philippe Gambette. Normalisation of 16^{th} and 17^{th} century texts in French and geographical named entity recognition. In *ACM SIGSPATIAL GeoHumanities'20*, Proceedings of the 4^{th} ACM SIGSPATIAL Workshop on Geospatial Humanities, pages 28–34, Seattle (virtual), United States, November 2020. ACM. https://hal-upec-upem.archives-ouvertes.fr/hal-02955867. 51

[KKL09] Orgad Keller, Tsvi Kopelowitz, and Moshe Lewenstein. On the longest common parameterized subsequence. *Theoretical Computer Science*, 410(51):5347–5353, 2009. http://doi.org/10.1016/j.tcs.2009.09.011. 56, 57

[KNTX08] Iyad A. Kanj, Luay Nakhleh, Cuong Than, and Ge Xia. Seeing the trees and their branches in the network is hard. *Theoretical Computer Science*, 401:153–164, 2008. https://doi.org/10.1016/j.tcs.2008.04.019. 14

[KPKW22] Sungsik Kong, Joan Carles Pons, Laura Kubatko, and Kristina Wicke. Classes of explicit phylogenetic networks and their biological and mathematical significance. *Journal of Mathematical Biology*, 84(6):47, 2022. http://doi.org/10.1007/s00285-022-01746-y. 11

[LBP10] Daniel Le Berre and Anne Parrain. The Sat4j library, release 2.2. *Journal on Satisfiability, Boolean Modeling and Computation*, 7(2-3):59–64, 2010. https://doi.org/10.3233/SAT190075. 16

[LCL18] Dominique Legallois, Thierry Charnois, and Meri Larjavaara. The balance between quantitative and qualitative literary stylistics: How the method of "motifs" can help. In Legallois, Charnois, and Larjavaara, editors, *The Grammar of Genres and styles*. De Gruyter Mouton, 2018. http://doi.org/10.1515/9783110595864-008. 35

[Leb16] Jean-Marc Leblanc. *Analyses lexicométriques des vœux présidentiels*. ISTE Group, 2016. 37

[LGMT10] Hyeran Lee, Philippe Gambette, Elsa Maillé, and Constance Thuillier. Densidées : calcul automatique de la densité des idées dans un corpus oral. In *RECITAL'2010 : 12ième Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 1–10, Montréal, Canada, July 2010. https://halshs.archives-ouvertes.fr/halshs-00495768. 28

[LL13] Cyril Labbé and Dominique Labbé. Existe-t-il un genre épistolaire? Hugo, Flaubert et Maupassant. *Nouvelles Journées de l'ERLA*, pages 53–85, 2013. https://hal.science/halshs-00436351v1. 37

[LLS^{+}15] James A. Lake, Joseph Larsen, Brooke Sarna, Rafael R. de la Haba, Yiyi Pu, Hyun-Min Koo, Jun Zhao, and Janet S. Sinsheimer. Rings reconcile genotypic and phenotypic evolution within the proteobacteria. *Genome Biology and Evolution*, 7(12):3434–3442, 12 2015. http://doi.org/10.1093/gbe/evv221. 7

[LMF^{+}02] Cédric Lamalle, William Martinez, Serge Fleury, André Salem, B Fracchiolla, A Kuncova, and A Maisondieu. Lexico 3, outils de statistique textuelle. *Manuel d'utilisation. Université de la Sorbonne Nouvelle*, 2002. https://lexi-co.com/ressources/manuel-3.41.pdf. 7, 40

[LSK$^+$22] Sarah Lutteropp, Céline Scornavacca, Alexey M Kozlov, Benoit Morel, and Alexandros Stamatakis. NetRAX: accurate and fast maximum likelihood phylogenetic network inference. *Bioinformatics*, 38(15):3725–3733, 2022. https://pubmed.ncbi.nlm.nih.gov/35713506/. 20

[Mar17] Claude Martineau. TreeCloud & Unitex: an increased synergy. ECLAVIT Workshop, November 2017. Poster, https://hal.science/hal-01702091. 33

[MMOS$^+$20] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.645. 41

[Moi20] Hermann Moisl. How to visualize high-dimensional data: a roadmap. *Journal of Data Mining & Digital Humanities*, 2020. https://jdmdh.episciences.org/7021/pdf. 37

[Mor05] Franco Moretti. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, 2005. 2

[Nak04] Luay Nakhleh. *Phylogenetic networks*. PhD thesis, The University of Texas at Austin, 2004. http://hdl.handle.net/2152/2126. 17

[NJZMC05] Luay Nakhleh, Guohua Jin, Fengmei Zhao, and John Mellor-Crummey. Reconstructing phylogenetic networks using maximum parsimony. In *2005 IEEE Computational Systems Bioinformatics Conference (CSB'05)*, pages 93–102. IEEE, 2005. https://doi.org/10.1109/csb.2005.47. 19

[Pel95] Jean-Christophe Pellat. Norme et variation orthographique au XVII$^{\text{e}}$ siècle. In *Rencontres linguistiques en pays rhénan 5/6*, Scolia : Sciences Cognitives, Linguistiques & Intelligence Artificielle, pages 245–260. Université des sciences humaines Strasbourg, 1995. https://www.persee.fr/doc/scoli_1253-9708_1995_act_3_1_889. 51

[PM16] Joan Carles Pons Mayol. *Reconstruction problems for LGT networks*. PhD thesis, Universitat de les Illes Balears, 2016. http://hdl.handle.net/11201/148945. 15

[Pos74] Rebecca Posner. Solomon Marcus, Poetica matematicä. Bucharest: Editura Academiei Republicii Socialiste România, 1970. pp. 400. (abstract and table of contents in English). *Journal of Linguistics*, 10(1):216–217, 1974. https://doi.org/10.1017/S0022226700004205. 55

[Rob51] William S. Robinson. A method for chronologically ordering archaeological deposits. *American Antiquity*, 16(4):293–301, 1951. http://www.jstor.org/stable/276978. 35

[SF09] Robert Sedgewick and Philippe Flajolet. *Analytic combinatorics*. Cambridge University Press, 2009. https://ac.cs.princeton.edu/home/. 22

[SGB15] Zied Sellami, Jean-Gabriel Ganascia, and Mohamed Amine Boukhaled. Medite: logiciel d'alignement de textes pour l'étude de la génétique textuelle. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Démonstrations*, pages 1–2, 2015. https://aclanthology.org/2015.jeptalnrecital-demonstration.1/. 48

[SGC+13] Lionel Spinelli, Philippe Gambette, Charles E. Chapple, Benoît Robisson, Anaïs Baudot, Henri Garreta, Laurent Tichit, Alain Guénoche, and Christine Brun. Clust&See: A Cytoscape plugin for the identification, visualization and manipulation of network clusters. *BioSystems*, 113(2):91–93, 2013. https://hal-upec-upem.archives-ouvertes.fr/hal-00832028. 27

[SGLP21a] Olga Seminck, Philippe Gambette, Dominique Legallois, and Thierry Poibeau. The corpus for idiolectal research (CIDRE). European Association of Digital Humanities Conference (EADH 2021), September 2021. Poster, https://hal.archives-ouvertes.fr/hal-03353520. 39

[SGLP21b] Olga Seminck, Philippe Gambette, Dominique Legallois, and Thierry Poibeau. The corpus for idiolectal research (CIDRE). *Journal of Open Humanities Data*, 7:15, 2021. http://doi.org/10.5334/johd.42. 39

[SGLP22] Olga Seminck, Philippe Gambette, Dominique Legallois, and Thierry Poibeau. The evolution of the idiolect over the lifetime: A quantitative and qualitative study of french 19th century literature. *Journal of Cultural Analytics*, 7(3), 2022. http://doi.org/10.22148/001c.37588. 34, 35, 39

[Slo94] Neil J. A. Sloane. An on-line version of the encyclopedia of integer sequences. *The Electronic Journal of Combinatorics*, pages F1.1–5, 1994. https://doi.org/10.37236/1194, OEIS available at http://oeis.org. 23

[SN87] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987. https://doi.org/10.1093/oxfordjournals.molbev.a040454. 30

[SS06] Charles Semple and Mike Steel. Unicyclic networks: compatibility and enumeration. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3:398–401, 2006. http://www.math.canterbury.ac.nz/~c.semple/papers/SS06.pdf. 21

[TGG11] Laurent Tichit, Philippe Gambette, and Alain Guénoche. ModClust: a Cytoscape plugin for modularity-based clustering of networks. In *MARAMI 2011*, Grenoble, France, October 2011. https://hal.archives-ouvertes.fr/hal-01261856. 27

[Vac10] Claire Hélène Vachon. *Le Changement linguistique au XVIe siècle : une étude basée sur des textes littéraires français*. ELiPhi, Éditions de linguistique et de philologie, 2010. 51

[Val02] Gabriel Valiente. *Algorithms on trees and graphs*, volume 112. Springer, 2002. https://link.springer.com/book/10.1007/978-3-662-04921-1. 2

[Val18] Edith Vallée. *Le Matrimoine de Paris : 20 itinéraires 20 arrondissements*. Christine Bonneton, 2018. 42

[VASJG10] Balaji Venkatachalam, Jim Apple, Katherine St. John, and Daniel Gusfield. Untangling tanglegrams: Comparing trees by their drawings. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(4):588–597, 2010. http://doi.org/10.1109/TCBB.2010.57. 36

[VIKS^{+}18] Leo Van Iersel, Steven Kelk, Georgios Stamoulis, Leen Stougie, and Olivier Boes. On unrooted and root-uncertain variants of several well-known phylogenetic network problems. *Algorithmica*, 80:2993–3022, 2018. https://doi.org/10.1007/s00453-017-0366-5. 14

[vISS10] Leo van Iersel, Charles Semple, and Mike Steel. Locating a tree in a phylogenetic network. *Information Processing Letters*, 110(23):1037–1043, 2010. https://doi.org/10.1016/j.ipl.2010.07.027. 14

[Wel18] Mathias Weller. Linear-time tree containment in phylogenetic networks. In *RECOMB-CG 2018*, pages 309–322, 2018. https://hal.science/hal-01802821v1. 15

[Zha16] Louxin Zhang. On tree-based phylogenetic networks. *Journal of Computational Biology*, 23(7):553–565, 2016. https://doi.org/10.1089/cmb.2015.0228. 17

Abstract

This habilitation thesis gathers several studies in computer science on the notions of proximity, similarity and heredity, with applications in bioinformatics and in digital humanities. The first part covers phylogenetic networks, which, reconstructed from similarities between biological data, make it possible to model complex hereditary relationships between biological species. Algorithmic solutions to reconstruction and characterization problems are presented, along with properties related to the counting of these networks or their components. The second part proposes several proximity-based data mining approaches that can be mobilized in digital humanities, in particular through network partitioning, word tree construction or the detection of temporal proximities in text corpora. The third part presents several methods for analyzing the heredity relationships between several texts, by aligning them at different scales, detecting intertextual relationships, or studying several versions at different successive states of the language. This work illustrates not only the possibility of adapting to digital humanities methods and algorithms inspired by bioinformatics, but also the interest of using simple digital models such as sequences, trees or networks to propose new methodologies for textual data analysis.

Résumé

Ce mémoire d'habilitation à diriger des recherches réunit plusieurs travaux en informatique sur les notions de proximité, de similarité et d'hérédité, avec des applications en bioinformatique et en humanités numériques. La première partie est dédiée aux réseaux phylogénétiques, qui, reconstruits à partir de similarités entre données biologiques, permettent de modéliser des relations d'hérédité complexes entre espèces biologiques. Des solutions algorithmiques à des problèmes de reconstruction et de caractérisation et des propriétés liées au comptage de ces réseaux ou de leurs composants sont présentées. La deuxième partie propose plusieurs approches d'exploration de données fondées sur leur proximité qui peuvent être mobilisées en humanités numériques, en particulier par du partitionnement de réseaux, la construction d'arbres de mots ou la détection de proximités temporelles dans des corpus de textes. La troisième partie présente plusieurs méthodes visant à analyser les relations d'hérédité entre plusieurs textes, qu'il s'agisse de les aligner à diverses échelles, d'y détecter des relations d'intertextualité, ou d'en étudier plusieurs versions à divers états successifs de la langue. Ce travail illustre non seulement la possibilité d'adapter aux humanités numériques des méthodes et algorithmes inspirés par la bioinformatique, mais aussi l'intérêt de se fonder sur des modèles informatiques simples de séquences, d'arbres ou de réseaux, pour proposer de nouvelles méthodologies d'analyse ou de visualisation de données textuelles.