Statistica Applicata - Italian Journal of Applied Statistics Vol. 35 (3)
doi.org/10.26398/IJAS.0035-016

VISUALIZATION OF TEXTUAL DATA: A COMPLEMENT TO AUTHORSHIP ATTRIBUTION

Ludovic Lebart[1]
Centre National de la Recherche Scientifique (CNRS), Paris, France

Abstract. *In textual data analysis, authorship attribution is precisely a leading case of statistical decision. While analyzing a large corpus of 50 French novels of the 20th century, we investigate the frontiers between descriptive (or unsupervised) methods, and confirmatory (or supervised) methods. It will be shown that additive trees applied to the coordinates of a preliminary correspondence analysis (CA) can provide both a description and an help for a decision. Our results aim at showing the complementarity between exploratory techniques and AI. in that field.*

Keywords: *Textual data visualization, Authorship attribution, Additive trees, CA.*

1. INTRODUCTION

If artificial intelligence (AI) methods often give excellent results in terms of authorship attribution, the specialist of the concerned texts sometimes remains frustrated by the binary and blind nature of the decision. In the framework of a problem of (literary) matching (50 novels written by 24 authors) and in the spirit of "Deep learning" which introduces the "unsupervised" in AI, we will show that the joint use of correspondence analysis (CA) in a mixed supervised/unsupervised framework (technique of supplementary variables/ visualized regression) makes it possible to both obtain satisfactory results and understand the context of these results. It will also be recalled in passing that CA (like regression) is also a particular case of neural networks, and fully deserves to be included in the panoply of AI techniques.

[1] Ludovic Lebart, ludovic@lebart.fr

2. A PROBLEM OF LITERARY MATCHING: 50 novels/ 24 authors

Text The following analyzes relate to a large corpus consisting of 50 novels from 24 francophone writers, selected and provided by Etienne Brunet (Brunet *et al.* 2021). Corpus Size: 3,501,883 words (tokens) among which 82,914 distinct words (types) containing 31,503 hapaxes (words appearing once). The corresponding Type/Token ratio (TTR), a decreasing function of the size of the corpus, is: TTR = 0.024.

By "analysis" we mean here a sequence of correspondence analysis (CA) of lexical tables, followed by an additive tree (AT) computed on a subspace of CA principal axes. The CA phase involves the chi-square distance (and its property of distributional equivalence) and gives the possibility to select the dimension of the subspace, allowing for a regularization of the data (see for example: Author *et al.*, 1977, 1984; Author, 1992). Starting the processing with a principal axes analysis brings this approach within the framework of deep learning, which recommends preliminary structural analyzes and a possible regularization of the data. But the tools remain geometric and transparent.

Table 1: List of 25 "authors" with their two selected titles [and their corresponding symbols for figures 1 and 2] (* = Nobel prize)

Ajar: Gros-Câlin & La vie devant soi [aja.grosca & aja.viedev]
Aragon: Les Beaux Quartiers & Blanche ou l'oubli [ara.beauxq & ara.blanch]
Breton: Nadja & L'Amou Fou [bre.Nadja & bre.amour]
Camus*: L'étranger & La Chute [cam.etrang & cam.chute]
Colette: Sido & La Vagabonde [col.sido & col.vagabo]
Duras: Barrage au Pacifique & L'Amant [dur.barag & dur.amant]
Ernaux*: La Honte & Les Années. [ern.honte & ern.annees]
Gary1: La Promesse de l'Aube & Les Racines du Ciel [gar.promes & gar.racine]
Gary2: Clair de Femme & Au_delà de cette limite [gar.clair & gar.delali]
Gide*: La Symphonie Pastorale & L'Immoraliste [gide.sympho & gide.immora]
Giono: Le Grand Troupeau & Le Hussard sur le toit [gio.grand & gio.hussar]
Giraudoux: Simon le Pathétique & Bella [gir.Simon & gir.Bella]
Gracq: Le Rivage des Syrtes & Un Balcon en forêt [gra.rivage & gra.balcon]
Le Clézio*: Hasard & Le Désert [cle.hasard & cle.desert]
Malraux: L'Espoir & Les Conquérants [mal.espoir & mal.conque]
Mammeri: La Colline oubliée & La Traversée [mam.colli & mam.traver]
Mauriac*: Le Baiser... & Le Mystère Frontenac [mau.baiser & mau.myster]
Montherlant: Les Célibataires & Les Bestiaires [mon.celiba & mon.bestia]
Pérec: L'Homme qui dort & Les Choses [pere.hommed & per.choses]
Proust: Du côté de chez Swann & Le Temps retrouvé [pro.cote & pro.temps]
Queneau: Le Chiendent & Zazie dans le métro [que.chiend & que.zaziem]
Saint-Exupéry: Courrier Sud & Terre des Hommes [exu.courri & exu.terreh]

Tournier: Vendredi ou les limbes... & Eléazar [tou.vendre & tou.eleaza]Vian: L'Ecume des jours & L'Automne à Pékin [via.ecum & via.auto]
Yourcenar: Mémoires d'Hadrien & L'Oeuvre au noir [you.memoi & you.oeuvre]

Note that one author appears three times in that list. "Romain Gary" (Gary1, Gary2, Ajar). At the origin of a famous literary deception, Gary managed to win the most prestigious French literary prize twice (Prix Goncourt) by hiding behind the name of "Emile Ajar". The double presence of Gary (triple, with Ajar) in the list aims to analyze more finely this oddity. However, all the results presented here remain still valid without this over-representation.

3. BRIEF REMINDER ABOUT THE TOOLS

3.1 SUPERVISED AND UNSUPERVISED MODELS

Let us remind that the "unsupervised approach" (exploratory or descriptive) is the counterpart of the "supervised approach (confirmatory or explanatory approach). Factor analysis, PCA, CA and clustering are unsupervised whereas discriminant analysis or regression methods are supervised.

External validation is the standard procedure in the case of supervised learning. Once the model parameters are estimated (learning phase), external validation is used to evaluate the model (generalization phase), usually with cross validation methods. External validation occurs in the context of correspondence analysis in two practical circumstances:

a) when the data set may be divided into two or more parts, one part being used to estimate the model, the other part used to verify the suitability of this model,

b) when certain metadata or external information are available to supplement the description of items.

We assume that external information is in the form of "supplementary elements". Note that a statistical validation (mostly bootstrap) is the indispensable complement of these technique.

3.2 ADDITIVE TREES (AT): THE PHYLOGENETIC EXPLOSION

AT technique will be extensively and exclusively used in the paper. These trees were originally proposed by Buneman (1971), then studied by Sattah and Tverski (1977). The concept of hierarchy at the base of the ascending classification was to approximate the initial distances by an ultrametric distance. Additive trees are less demanding. More flexible than the Minimum Spanning Tree which depends on *n-1* parameter, the AT implies *2n-3* parameters. It remains to find an approximation of the initial distances which satisfies these conditions. With AT distance, a tree can be drawn with the objects as nodes, such that the distance between two objects is the length of the path joining these two objects on the tree.

Stimulated by the works of Barthélémy and Guénoche (1988), tree analysis methods have been widely used in the field of text analysis. However, the first proposed algorithms required a prohibitive computation volume for large numbers of objects to classify. Saitou and Nei (1987) proposed an algorithm called Neighbor Joining which approximately reduces the search for the additive tree to a classical ascending classification procedure. This heuristic which was implemented by Huson and Bryant (2006) [SplitsTree] and used here, had a huge impact on the rapidly expanding world of phylogenetic research. Saitou and Nei's article has been cited more than 68,000 times since its publication. Theoretical justifications for the algorithm's efficiency were presented by Mihaescu *et al.* (2009).

3.3 CORRESPONDENCE ANALYSIS AS A NEURAL NETWORK

The links between Singular Value Decomposition (SVD) and Principal Components Analysis (PCA) with some particular neural networks have been stressed by Bourlard and Kamp (1988), Baldi and Hornik (1989), Asoh and Otsu (1989). Correspondence Analysis (Benzécri, 1969; or its non-symmetrical version, Lauro and D'Ambra, 1984; Balbi, 1994; Balbi and Triunfo, 2013) is at the meeting point of many techniques. It can be described as both supervised and unsupervised multilayer perceptrons (Author, 1997). In the supervised case, the input and the output layers are respectively the rows and the columns of the contingency table. In the unsupervised case, both the input layer and the output

layer could be the rows, whereas the observations could be the columns of the table. In both situations, the networks make use of the identity function as a transfer function. More general transfer functions might lead to interesting non-linear extensions of the method.

3.4 SUPPLEMENTARY VARIABLES AND REGRESSION

Adding supplementary elements in a principal axes technique (SVD, PCA, CA) constitutes a descriptive variant of the multiple regression (being itself a simple form of perceptron). From a geometrical point of view, the two situations are indeed similar (see, e.g.: Lebart *et al.*, 1984, 2019):

Regression: The p explanatory variables generate a subspace having at most p dimensions on which is projected the variable y to explain.

CA or PCA: the p active variables of the analysis also generate a subspace with at most p dimensions that we reduce to q factors to visualize it. It is on this subspace reduced to q dimensions that we project afterwards the supplementary variables to locate them with respect to the active variables. A visualization is then possible in the space spanned by every pair of axes.

All the following results have been obtained with the help of the freely downloadable software DtmVic (www.dtmvic.com).

4. MAIN RESULTS

The analysis will focus on vocabulary, more in the spirit of content analysis than in the context of stylometry. We do not seek to discriminate between authors, and the pairings of texts observed in the forthcoming graphical displays will come somewhat as a surprise, a "statistical fact". We will lemmatize the corpus, using the free software *TreeTagger* (Schmid, 1994) discarding function words, proper nouns or personal pronouns (to eliminate, for instance, the effects of narrative at first person) (Section 4.1). Then we study the subset of verbs alone (Section 4.2). Finally, Section 4.3 analyzes directly and blindly the pages (here: sequences of 50 lines), without reference to a novel (the novels being positioned *a posteriori* as centroids of their pages). The approach followed is then similar, in a descriptive framework, to the so-called *Word2vec* approaches of AI.

4.1 BASIC GLOBAL ANALYSES ON LEMMAS

Figure 1 displays the first visualization of the whole lemmatized corpus, for a frequency threshold of 100 (2364 lemmas). The results are satisfactory: only one author escapes the matchings: Montherlant. The divergence between his two novels “Bestiaires” and “Célibataires” will be the big exception for many approaches. These two novels are indeed from the same author, but they call upon pools of exceptionally different vocabularies for one and the same author: we will see later that this difference is detectable even on verbs alone. Note that the tree of figure 1 remains similar with identical conclusions for a smaller frequency threshold of 50 corresponding to 4018 lemmas.

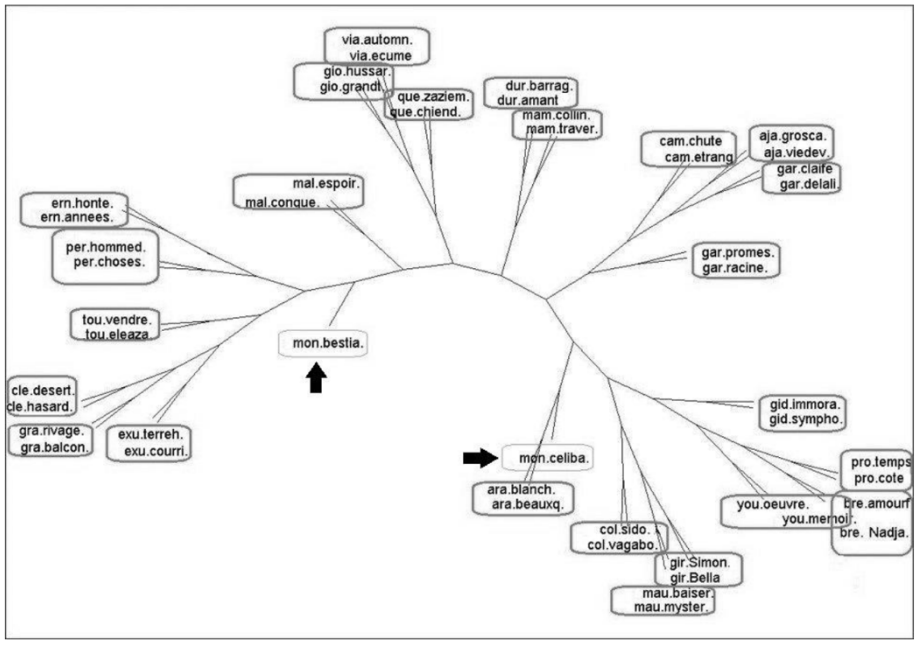

Figure 1: Additive tree for 50 novels de scribed by 2364 words (lemmas) appearing at least 100 times in the corpus. All CA axes (49) are kept. Only one author corresponds to unmatched novels (black arrows): Montherlant (novels: *Bestiaires* and *Les Célibataires*).

4.2 LIMITING THE VOCABULARY TO VERBS

The second approach concerns only verbs. Verbs are much less characteristic of a specific novel than nouns or adjectives, obviously more linked to the content of the text. Figure 2, however, gives us a surprising good result: 23 authors have been correctly matched (except Camus and again Montherlant).

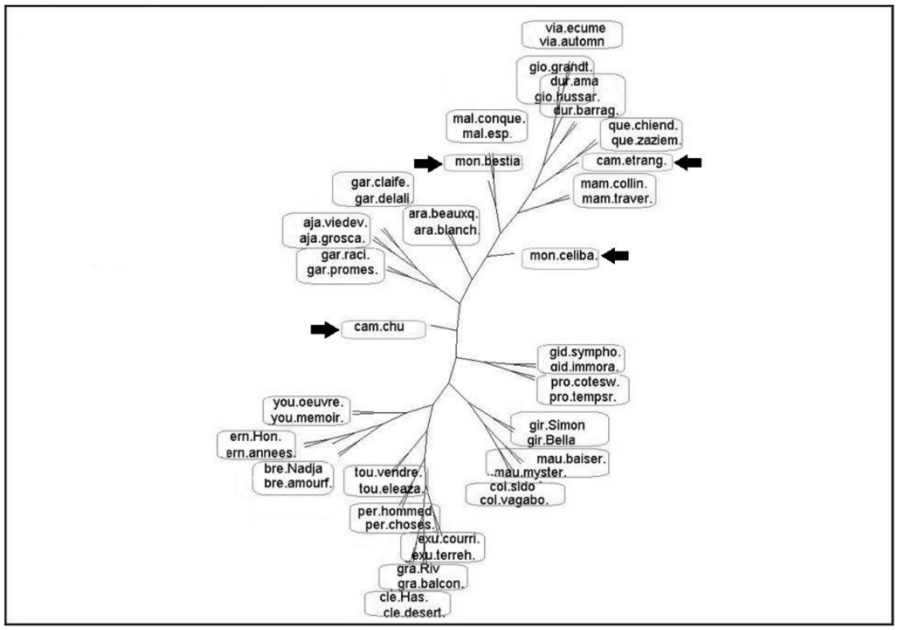

Figure 2: Additive tree for the 50 novels described only by their 726 verbs (without auxiliary verbs such as "to be", "to have". Frequency threshold for verbs: 84). Misclassified: Camus, Montherlant (black arrows).

4.3 FRAGMENTED NOVELS: ANALYSING THE 3547 PAGES

Finally, the third approach presented here is radically different. This time, the basic analysis is completely unsupervised. New "artificial observations" can be created in a text corpus, generalizing to large fragments the *context units* of the original approach proposed by Reinert (1983) at the basis of a procedure known as ALCESTE methodology.

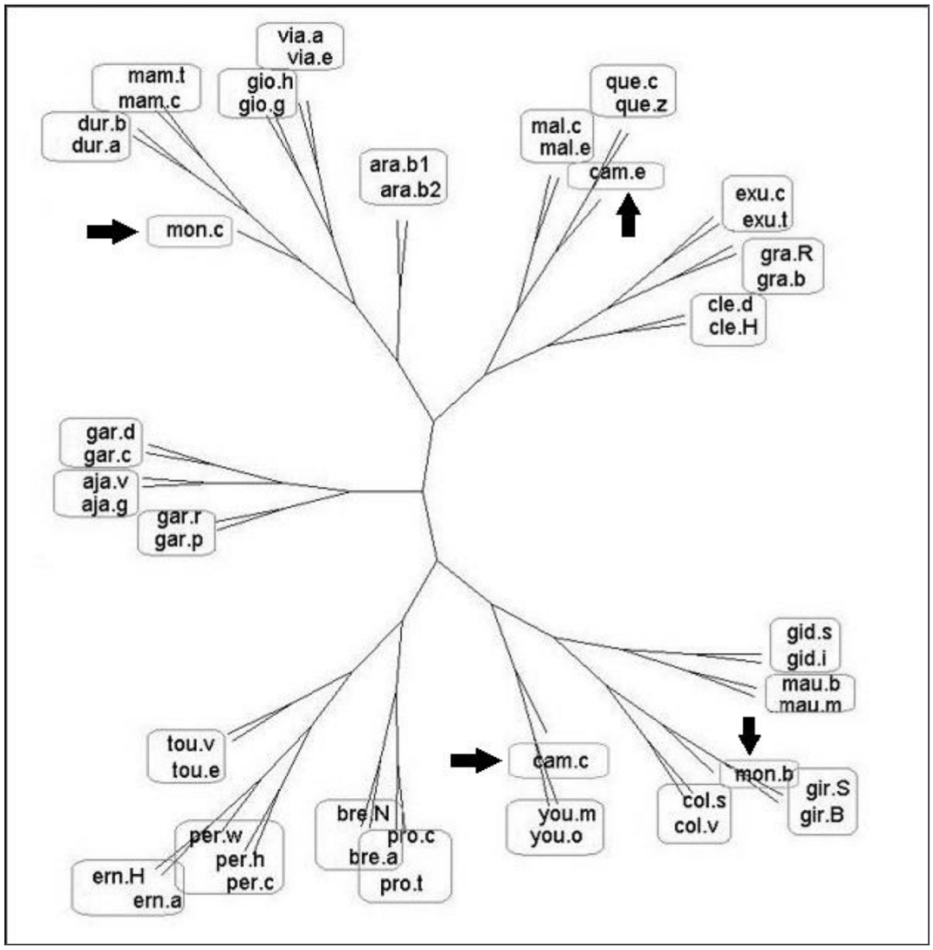

Figure 3: Verbs only. Additive tree built from the coordinates of a totally unsupervised CA of 3547 pages of 50 lines. Frequency threshold for the 726 verbs: 84. *A posteriori* projections of the centroids of pages belonging to a same novel. Misclassified: 2 authors out of 25: Camus, Montherlant (black arrows).

The advantages of the fragmentation of the corpus are the following:

- The structure of the text **inside** each novel is now taken into account, a piece of information overlooked in the classical approach to the single aggregate table of Sections 4.1 and 4.2. This entails a deeper understanding of the internal structure of each text, a finer granularity.

- An external validation evidence can then be achieved using the partition of the initial corpus of texts (the classes of which being the novels).

We are now dealing with an analysis of the 3547 pages of 50 lines (which rather correspond to printed double pages) of the lemmatized file, which can be shuffled like playing cards. Once the typology of these pages has been obtained, the novels are positioned as the average points of their pages. The analysis does not seek to contrast the novels, but to contrast the pages. It is therefore a very severe test.

If we fragment into pages the lemmatized corpus of Section 4.1, 17 authors (out of 25) are well matched. That result will be improved by the fragmentation into pages of the corpus limited to verbs used in Section 4.2.

We mentioned above that verbs were more evenly distributed in the texts than nouns and adjectives. We will now continue working on verbs (pages of verbs) (Figure 3) to observe that verb pages allow better prediction than word pages (lemmas). Indeed, despite the severity of the test, 23 out of 25 authors are characterized by their pages of verbs.

Only the two writers Montherlant and Camus are left to make an exception to this new endeavor to match novels. Note that “Bestiaires” is an autobiographic novel written by the young Montherlant passionate with bullfighting, whereas the second novel “Les célibataires”, is dedicated to the sad end of life of two elderly bachelors. For Camus, “L’étranger” is his first novel, and “La chute” his last one.

Evidently, these remarks inspired by external pieces of information are only sketches and hypotheses that can be improved and enriched with all the available tools and parameters of these exploratory phases: levels of fragmentation (paragraphs, pages, chapters), grammatical units (function words, nouns, adjectives), size of the subspace of coordinates, thresholds of frequencies for lexical units.

CONCLUSIONS

At this stage, we have combined several techniques. Regularization through Principal Axes Techniques (here: CA), fragmentation (related or similar to *Word2vec* approach), projection of supplementary (or illustrative) variables, nearest neighbors prediction (through Additive trees representation) that can be expressed either in terms of Neural Networks and Machine Learning, or more aptly in terms of deep learning (Vanni *et al.*, 2018).

About Deep Learning, let us quote the inspiring remark of Le Cun, Bengio and Hinton (Le Cun et al., 2015): "…we expect unsupervised learning to become far more important in the longer term. Human and animal learning is largely unsupervised: we discover the structure of the world by observing it, not by being told the name of every object".

In the field of textual data analysis, the priority is not systematically "recognition" but discovery, description, comparison, understanding. Such approach remains partially supervised in the sense that both the available external information and the discovered structures are used to enhance the exploration.

But within Machine Learning toolbox, we have selected transparent procedures, interpretable at each step, whose results could be either visualized (planes, trees), or assessed via statistical procedures (bootstrap). Obviously, the selected methods are only a part of the potential of machine learning. But this was the price to pay for the transparency and the algebraic simplicity of the process. Using the arsenal of black boxes available, the machine learns. Using the subset of selected visualization techniques, the researcher learns, we learn.

REFERENCES

Asoh, H. and Otsu, N. (1989): Nonlinear Data Analysis and Multilayer Perceptrons. IEEE, IJCNN-89, 2, 411-415.

Balbi, S. (1994). L'Analisi Multidimensionale dei dati negli anni'90. Dipartimento di Matematica e Statistica. (Univ. Federico II), Rocco Curto Editore, Napoli.

Balbi, S. and Triunfo, N. (2013). Statistical tools in the joint analysis of closed and open-ended questions. In: Survey data Collection and Integration. Davino C., Fabbris L. (eds). Springer Verlag, Berlin. 61-74.

Baldi, P. and Hornik, K. (1989): Neural networks and principal component analysis: learning from examples without local minima. Neural Networks, 2: 52-58.

Barthélémy, J.-P. and Guénoche, A. (1988). *Les arbres et les représentations de proximité*. Masson, Paris.

Benzécri, J.-P. (1969): Statistical analysis as a tool to make patterns emerge from clouds. In: *Methodology of Pattern Recognition*, S. Watanabe, (ed.) Academic Press: 35-74.

Bourlard, H. and Kamp, Y. (1988): Auto-association by Multilayers perceptrons and singular value decomposition. *Biological Cybernetics*, 59: 291-294.

Brunet, E., Lebart., and Vanni, L. (2021) Littérature et intelligence artificielle. In: Mayaffre D., Vanni L., (eds) *L'Intelligence Artificielle des Textes*. Honoré Champion, Paris : 73-128.

Buneman, P. (1971). The recovery of trees from measurements of dissimilarity. In: Hodson F. R. D. Kendall G., and Tautu P., (Editors). *Mathematics in the Archeological and Historical Sciences*. Edinburgh University Press: 387-395.

Huson, D.H. and Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, *Molecular Biology and Evolution*, vol. (23), 2: 254-267.

Lauro, N. C and D'Ambra, L. (1984): L'Analyse non-symétrique des Correspondances. In: *Data Analysis and Informatics*, III, Diday *et al.* (eds.), North-Holland: 433-446.

Le Cun, Y., Bengio, Y. and Hinton G. (2015). Deep Learning, *Nature*, 521, 436-444.

Lebart, L., Morineau, A. and Tabard N. (1977). *Technique de la Description Statistique*. Dunod, Paris.

Lebart, L., Morineau A. and Warwick K. (1984). *Multivariate Descriptive Statistical Analysis*. John Wiley and Sons, New York.

Lebart, L. (1992). Discrimination through the Regularized Nearest Cluster Method. In: *Computational Statistics*, Y. Dodge *et al*. (eds.), Springer Verlag, Berlin, Heidelberg: 103-118.

Lebart, L. (1997). Correspondence analysis, discrimination and neural networks. In: *Data Science, Classification and Related Methods*. Hayashi C., Ohsumi N., Yajima K., Tanaka Y., Bock H.- H. and Baba Y. (eds), Springer, Berlin, 423-430.

Lebart, L., Pincemin, B. and Poudat, C. (2019). *Analyse des Données Textuelles*, PUQ, Québec, Canada.

Luong, X. (1988). *Méthodes d'analyse arborée. Algorithmes, applications*. Thèse pour le doctorat ès sciences. Université Paris V.

Mihaescu, R., Levy, D. and Pachter, L. (2009). Why Neighbor-Joining works? *Algorithmica*, vol. (54): 1-24.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: Application à l'analyse lexicale par contexte. *Cahiers de l'Analyse des Données*, vol. (3): 187-198.

Saitou, N. and Nei, M. (1987). The neighbor joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, vol. (4), 4: 406-425.

Sattah, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika*, vol. (42), 3: 319-345.

Schmid, H. (1994). Probabilistic part of speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Vanni, L., Mayaffre, D. and Longrée, D. (2018). ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables, *JADT 2018, Universitalia,* Rome, ⟨hal-01823560⟩.