DSSR 2024
Data Science & Social Research
4th International Conference

Naples, 25th – 27th March 2024

Visualization of Textual Data,

Some recent improvements: simultaneous additive trees

Ludovic Lebart

CNRS (R),
ludovic@lebart.org

It is, in some respect, a continuation of :

ISA-RC33 7th International Conference on Social Science Methodology

Campo di Monte Sant’Angelo, Napoli, September 1 – 5, 2008.

Between principal axes analysis and clustering: the missing links.

Ludovic Lebart

Telecom-ParisTech,
46 rue Barrault, 75013, Paris, France
ludovic@lebart.org

With the involvement of Simona Balbi

Visualization of Textual Data,

Some recent improvements: simultaneous additive trees

Part 1. Visualization of data in Social Sciences

1.1 Basic tech. : Data compression; images, Graphs

Images (Example 1: Baalbek)
Graphs (Example 2: Map of Ireland)

1.2 Data Visualization problems in Social Sciences

Open questions (Example 3:Survey USA - Japan)
Semantic networks (Example 4: French verbs)

Part 2. Simultaneous Additive Trees

2.1 Additive trees (AT): the phylogenetic explosion
2.2 Simultaneous representation in CA *(reminder)*

2.3 Drawing simultaneous trees

(Example 5: Inaugural Address corpus)
(Example 6: Shakespeare Sonnets)
(Example 7: Georges Brassens)
(Example 8: Leonard Cohen)

Conclusion

Part 1. Visualization of data in Social Sciences

Clustering methods and **principal axes techniques** (principal components analysis, two-way and multiple correspondence analysis, canonical and linear discriminant analyses, etc.) have been often interacting during the last fifty years.

Most practitioners consider clustering methods and principal axes techniques (principal components analysis (PCA), correspondence analysis, (CA, MCA, etc.) as **complementary approaches** in the exploration of multivariate data sets.

As far as visualizations of data are concerned, the enrichment resulting from the simultaneous use of both families of methods is widely recognized.

A review of works at the intersection of these two fields of research reveals yet a wealth of algorithms often adapted to various empirical contexts.

At the outset, back to the first half of the twentieth century, the rotations using some specific criteria (involving some moments > 2) in the framework of factor analysis could be viewed as the first attempts to find clusters of variables.

Pragmatical standpoint

Several pragmatical methodologies, often present in Text Mining softwares, make use of both Clustering and Principal Axes Analyses techniques:

- 2.1 Clustering from principal coordinates,
- 2.2 *A posteriori* projection of clusters onto principal planes,
- 2.3 Dissection of a continuous space to automatically describe it : generalized histogram.
- 2.4 Use of the Minimum Spanning Tree to complement principal axes visualizations.
- 2.5 Use of Additive Trees (or Phylogenetic Trees) to get planar visualization summarizing more than 3 dimensions
-

A reminder: The first **Unsupervised approach**.

1904 : « a *[discrete]* breakthrough ».

Charles Spearman (1904) – "General intelligence, objectively determined and measured". *Amer. Journal of Psychology, 15, p 201-293.*

General factor for individual i

$$x_i^j = a_j f^i + \varepsilon_i^j$$

Value of variable j for individual i

Residual (hopefully small)

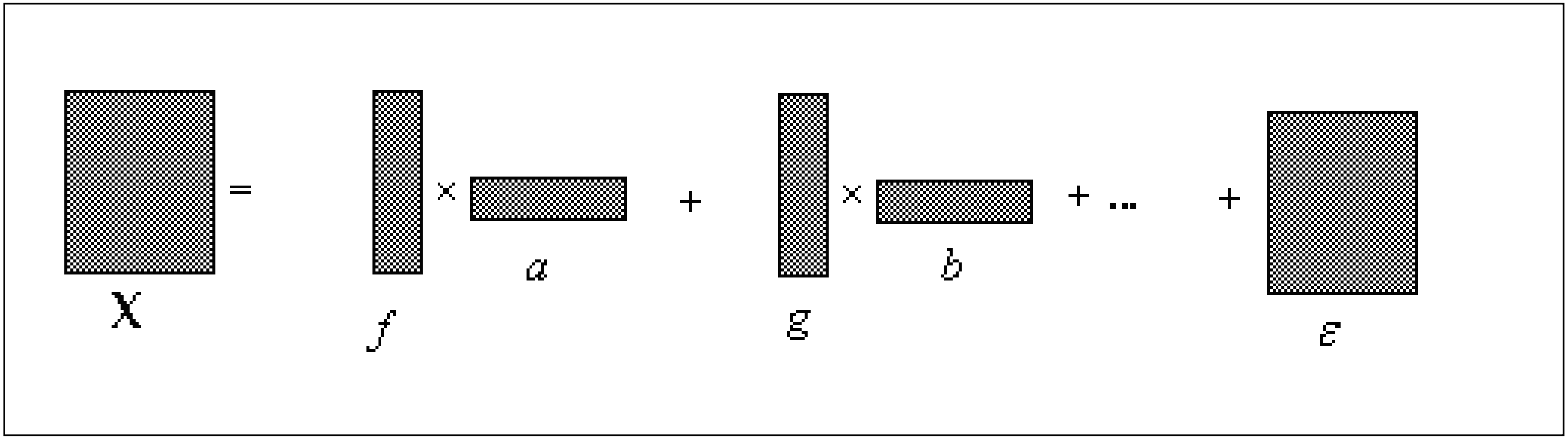
Coefficient of variable j

Known = ***Unknown***

... and its generalization to several factors

Garnett J.-C. (1919) - General ability, cleverness and purpose. *British J. of Psych.*, 9, p 345-366.
Thurstone L. L. (1947) - *Multiple Factor Analysis.* The University of Chicago Press, Chicago.

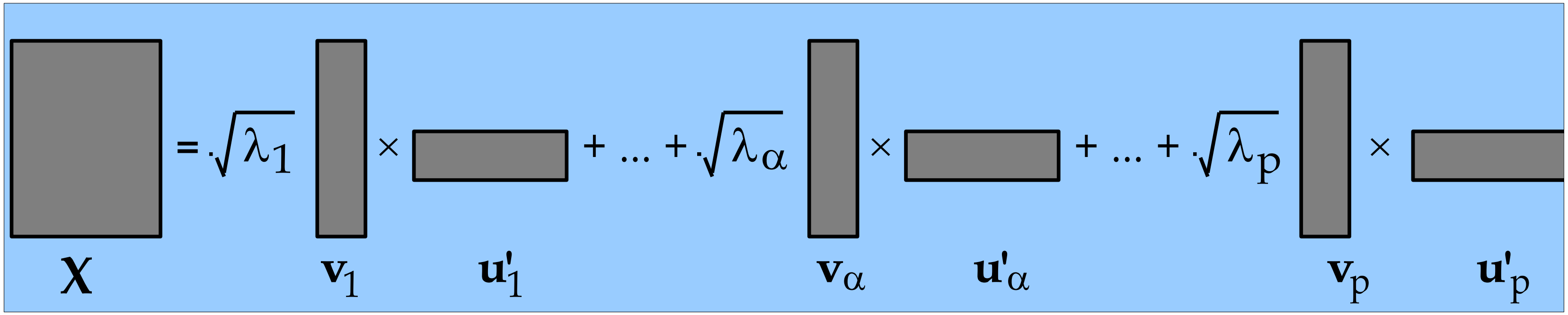
$$x_i^j = a_j f^i + b_j g^i + \ldots + \varepsilon_i^j$$



Singular Values Decomposition is a theorem, not a model

Eckart C., Young G. (1936) - The approximation of one matrix by another of lower rank. *Psychometrika*, I, p 211-218.

Eckart C., Young G. (1939) - A principal axis transformation for non- Hermitian matrices. *Bull. Amer. Math. Assoc.*, 45, p 118-121.



A precursor: Pearson K. (1901) - On lines and planes of closest fit to systems of points in space. *Phil. Mag.* 2, n°II, p 559-572.

Basic of Fourier Series: A multiple regression on orthogonal variables (functions of a single variable *t*).

$$a_0 = \frac{2}{T}\int_0^T f(t)dt$$

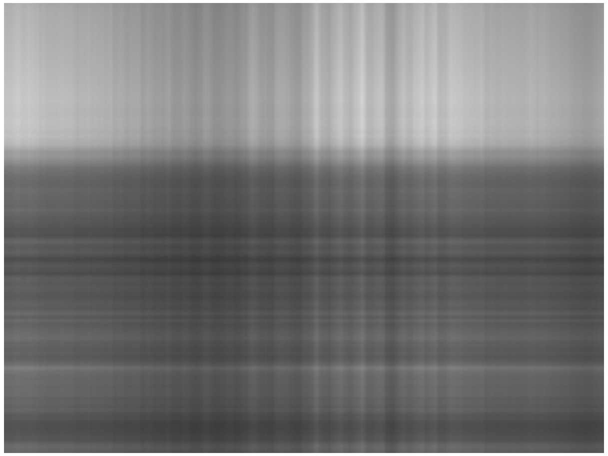
$$a_n = \frac{2}{T}\int_0^T f(t)\cos(\frac{2n\pi t}{T})dt$$

$$b_n = \frac{2}{T}\int_0^T f(t)\sin(\frac{2n\pi t}{T})dt$$

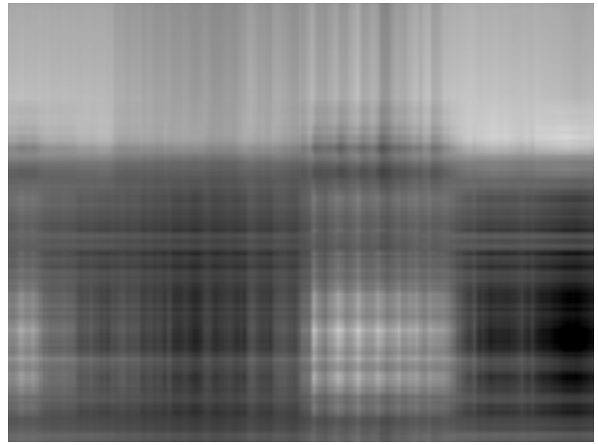
$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}[a_n \cos\frac{2n\pi t}{T} + b_n \sin\frac{2n\pi t}{T}]$$

Example 1
Baalbek Temple (Lebanon)

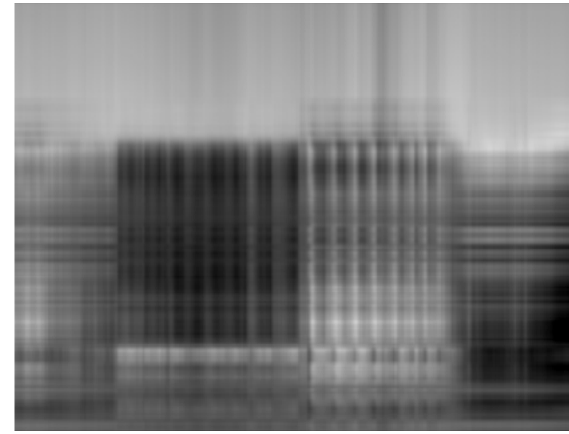
Example 1: Principal Axes Compression vs Fourier Compression



SVD 1 *(First term)*



SVD 2



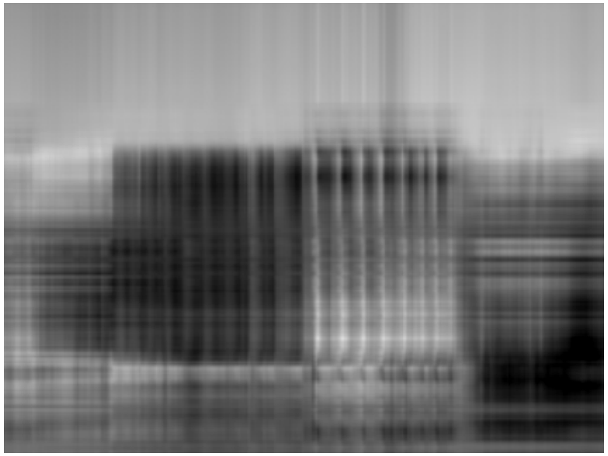
SVD 3

Fourier 1 *(First term)*

Fourier 2

Fourier 3

Example 1: Principal Axes Compression vs Fourier Compression (*continuation*)

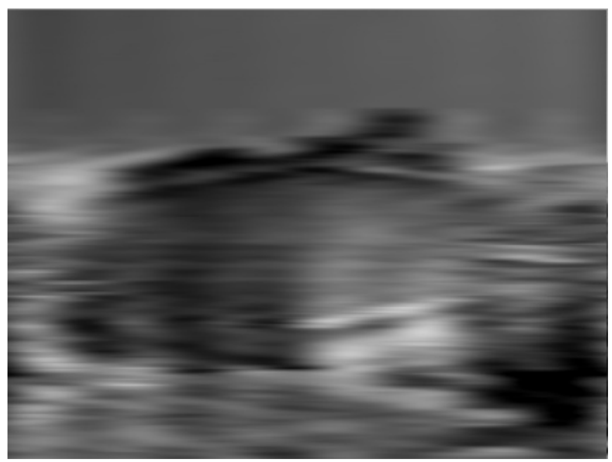


SVD 4

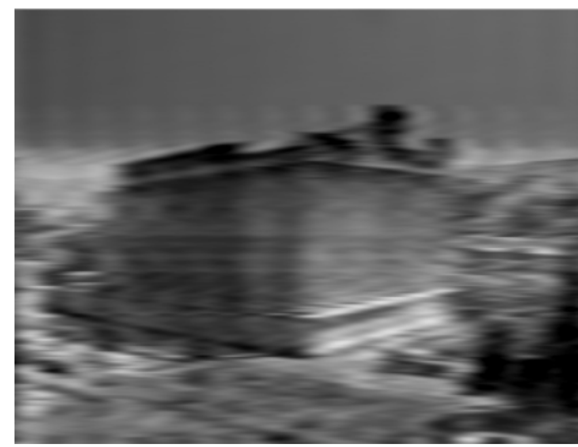
SVD 10

SVD 20

Fourier 4



Fourier 10



Fourier 20

Example 1: Principal Axes Compression vs Fourier Compression (*end*)

SVD 40

SVD 100

Fourier 40

Fourier 80

Fourier 160

A pedagogical example : Description of a textual graph

Each Irish county "answers" to the fictitious "open-question" :
Which are your neighboring counties?

Table 1: Text encoding contiguity relationship for four Irish counties

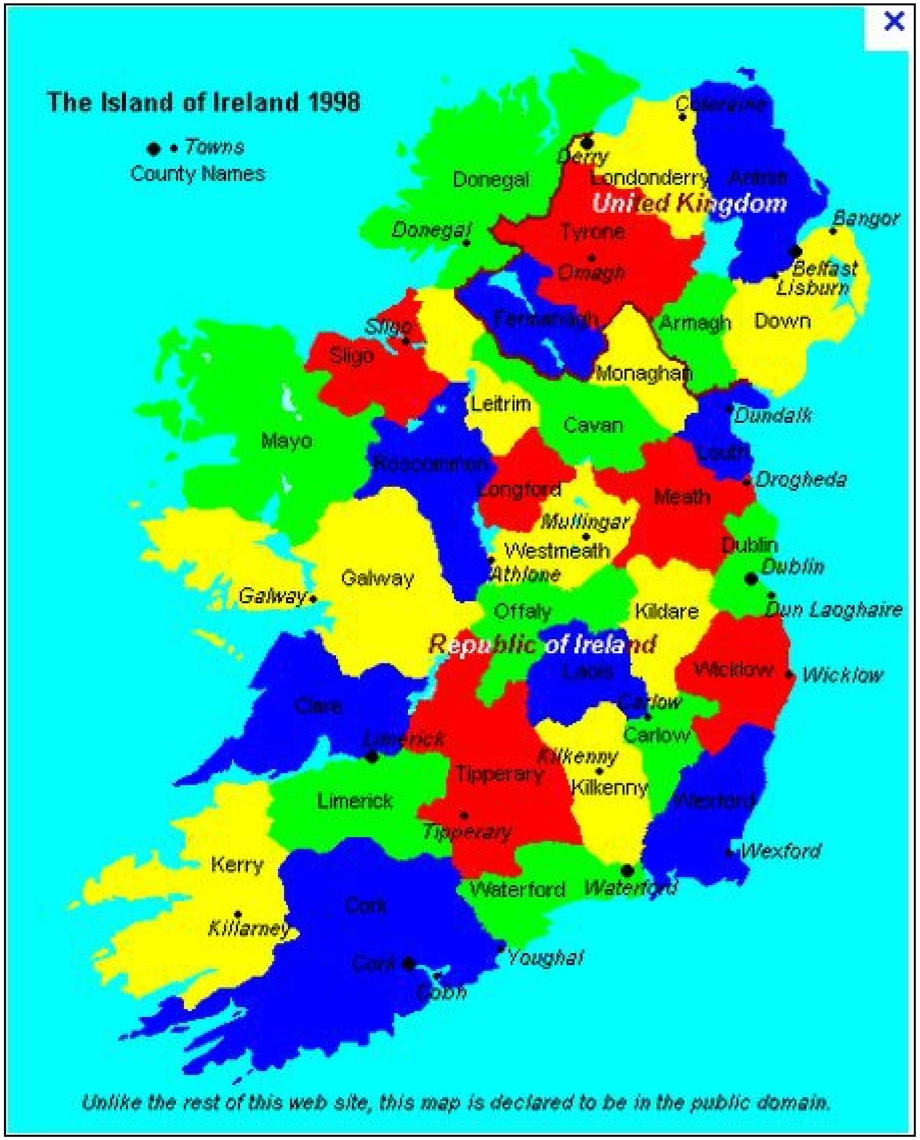
```
****     Galway
   Mayo  Roscommon  Offaly  Clare  Tipperary

****     Leitrim
   Sligo  Roscommon  Longford  Fermanagh  Cavan  Donegan

****     Mayo
  Sligo  Roscommon  Galway

****     Roscommon
   Sligo  Leitrim  Longford  Westmeath  Offaly
```
....

Example 2. Visualization of graphs (*continuation*)

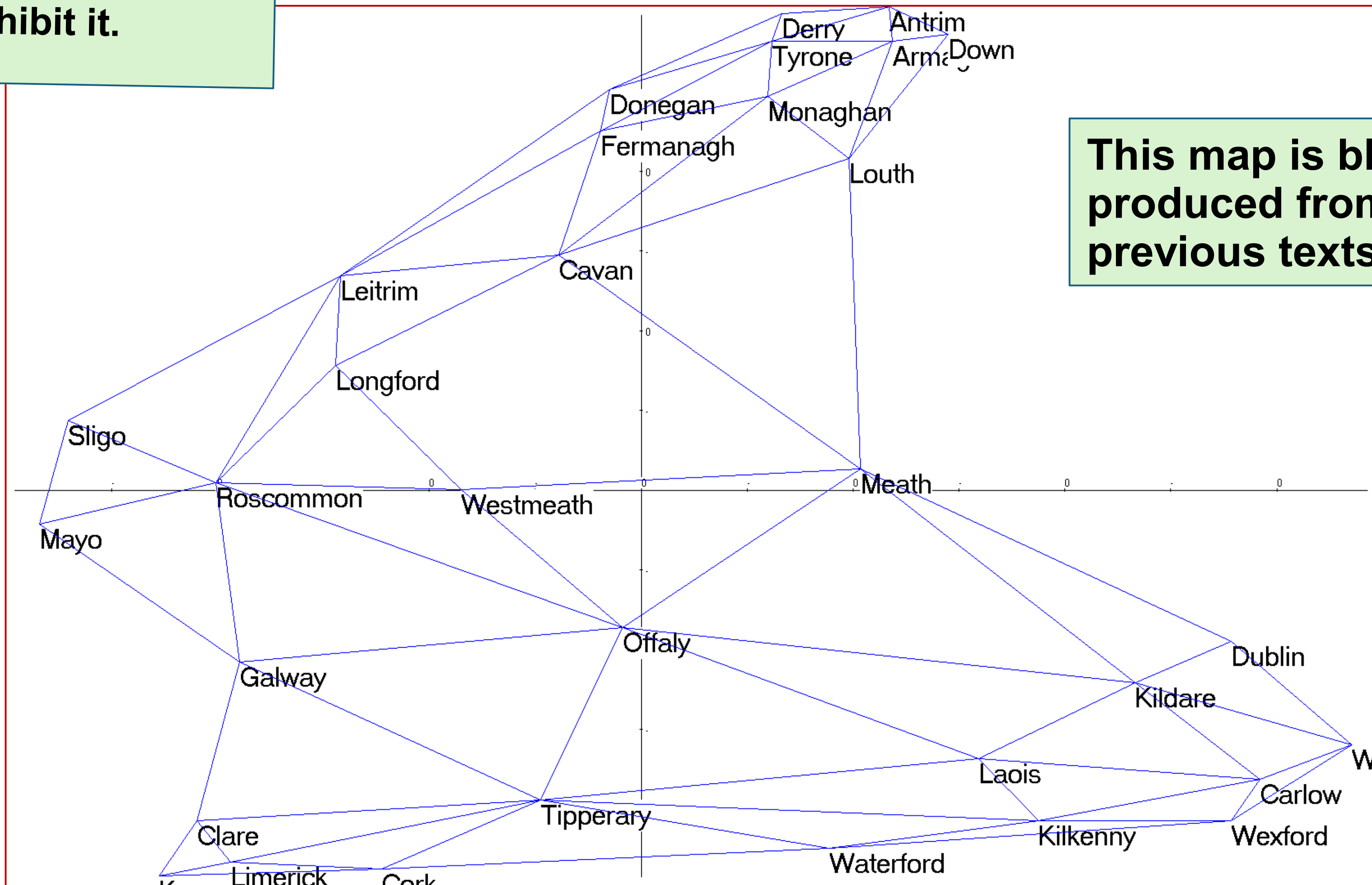


The capabilities of correspondence analysis to describe undirected graphs (of the flat grid or geographic map type) from their associated matrices were highlighted by Benzécri (1973, chap. 10) and Lebart *et al.* (1998).

Example 2. Visualization of graphs (*continuation*)

When a pattern exists within a text, some techniques may detect it and exhibit it.

Plane spaned by axes 1 and 2 of the CA of the lexical table (counties x counties)

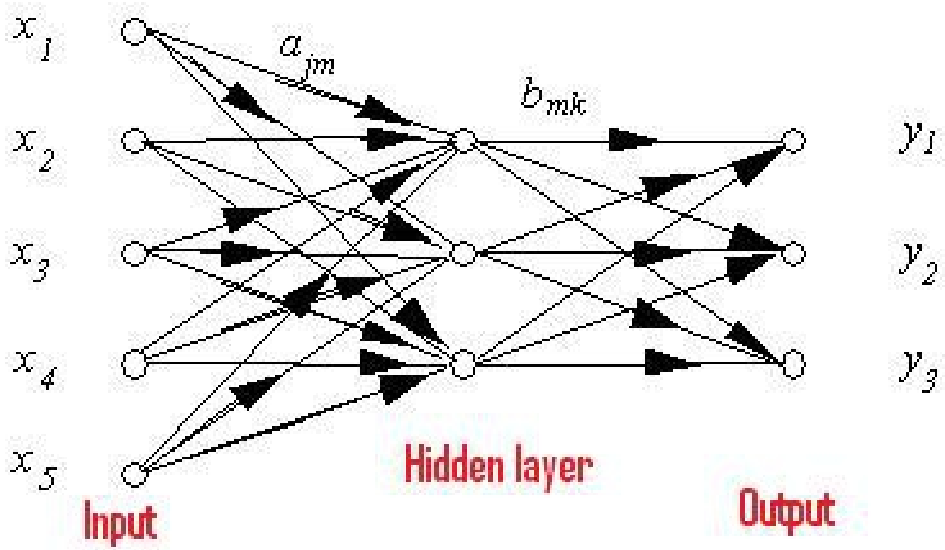


This map is blindly produced from the previous texts.

No « statistics » involved

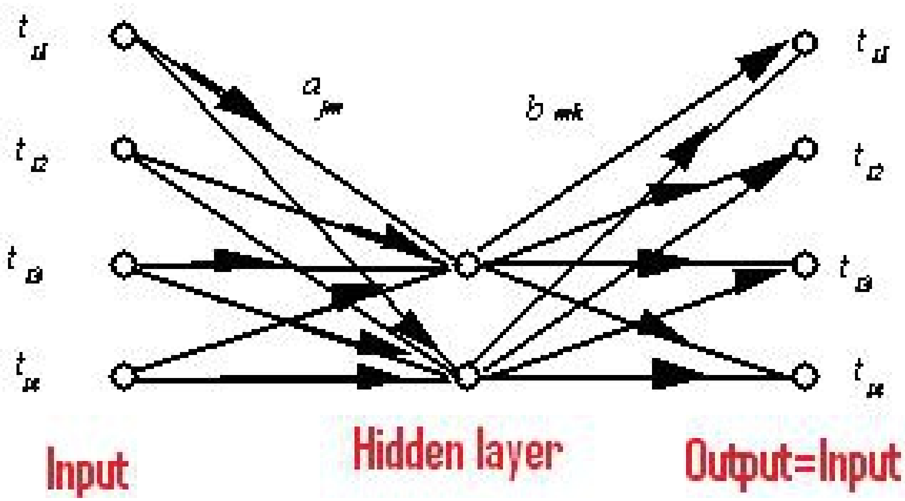
Incidentally: Principal Axes Compression and Neural networks [Perceptron]

Multi-layer Perceptron



$$y_{ik} = \Psi\left\{\sum_{m=1}^{r} b_{mk}\ \Phi\left(\sum_{j=1}^{q} a_{jm}x_{ij} + c_m\right) + d_k\right\} + e_{ik}$$

Self organised Perceptron



Supervised and unsupervised Models

In statistical learning theory:

"Unsupervised Approach" (exploratory or descriptive).
"Supervised approach (confirmatory or explanatory approach).

Factor analysis, PCA, CA and clustering are unsupervised,
Discriminant analysis, regression, neural networks methods are supervised.

The techniques of supplementary (or illustrative) variables could be considered as a bridge between supervised and unsupervised approaches.

External validation is the standard procedure in the case of supervised learning.

Once the model parameters were estimated (learning phase), external validation is used to evaluate the model (generalization phase), usually with validation methods such as cross-validation or/and bootstrap.

An international survey (Tokyo Gas Company)

A survey in three cities (Tokyo, New York, Paris) about dietary habits.

The 2 common <u>open-ended</u> questions were:

"*What dishes do you like and eat often?*
(With a probe: "*Any other dishes you like and eat often?*").

“ *What would be an ideal meal?”*

Akuto H.(Ed.) (1992). *International Comparison of Dietary Cultures*, Nihon Keizai Shimbun, Tokyo.

Akuto H., Lebart L. (1992). Le Repas Idéal. Analyse de Réponses Libres en Anglais, Français, Japonais. *Les Cahiers de l'Analyse des Données*, vol XVII, n°3, Dunod, Paris.

Example : An international survey (*continuation*)

"*What dishes do you like and eat often?*
"What would be an ideal meal?"
[Four responses (New York)]

---- 1
SPAGHETTI,CHINESE
++++
CAESAR SALAD,LOBSTER TAILS,BAKED POTATO, CHOCOLATE MOUSSE

---- 2
SEAFOOD,GREEN SALAD,CHINESE FOOD
++++
CHAMPAGNE,CAVIAR,GREEN SALAD,GRILLED SEAFOOD

---- 3
CHINESE FOOD
++++
CHINESE FOOD,FRENCH FOOD,VEAL,BREAD
---- 4
PASTA
++++
BEARNAISE BEEF,CHINESE FOOD,ITALIAN FOOD,PASTA

An international survey (Tokyo Gas Company)

The common open-ended question : "*What dishes do you like and eat often?*
(With a probe: "*Any other dishes you like and eat often?*").

- Sub-sample 1 (*city of Tokyo)* : 1008 individuals.
The global corpus of open responses contains 6219 occurrences of 832 distinct words. 139 words appear at least 7 times, leading to 4975 occurrences.

- Sub-sample 2 *(city of New-York)* contains 634 individuals.
(6511 occurrences of 638 distinct words).
The processing takes into account the 83 words appearing at least 12 times.

- Sub-sample 3 *(city of Paris)* contains 1000 individuals.
The global corpus contains 11108 occurrences of 1229 distinct words.
The processing takes into account the 112 words appearing at least 18 times, leading to 7806 occurrences.

- The three sets of respondents are broken down into into six categories (three categories of age, combined with the gender).

An international survey (Tokyo Gas Company)

words (frequency order)

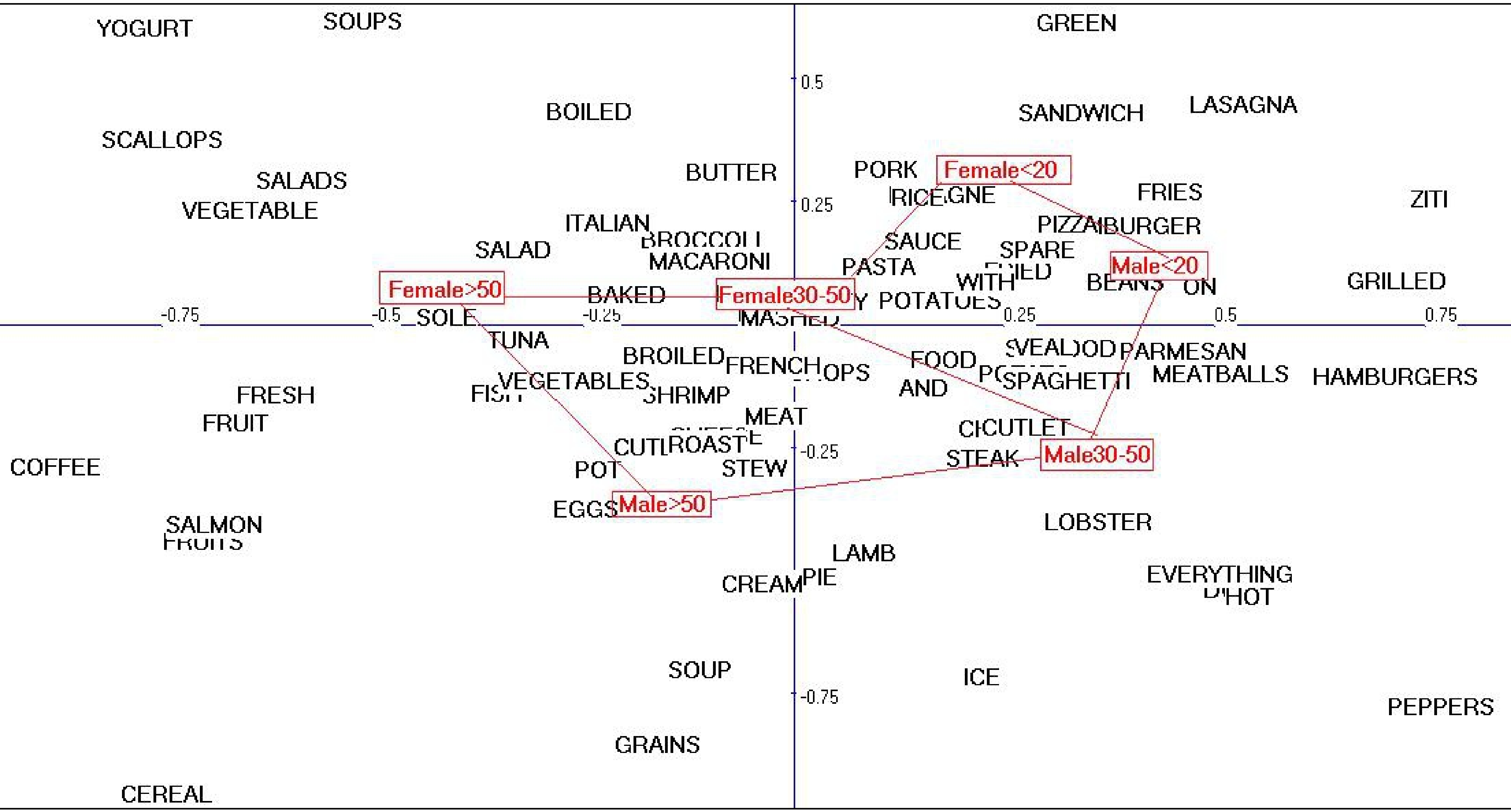
num.	used words	freq.
12	CHICKEN	254
73	STEAK	101
49	PASTA	95
22	FISH	87
60	SALAD	85
1	AND	85
23	FOOD	82
52	PIZZA	62
79	VEGETABLES	57
4	BEEF	56
71	SPAGHETTI	55
13	CHINESE	54
80	WITH	48
59	ROAST	47
58	RICE	45
67	SHRIMP	45
43	MACARONI	42
56	POTATOES	39
35	HAMBURGERS	36
75	TUNA	35
26	FRIED	33
77	VEAL	33
38	ITALIAN	31
2	BAKED	29
48	PARMESAN	29
55	POTATO	27
46	MEATBALLS	25
3	BEANS	24
45	MEAT	24
76	TURKEY	24
14	CHOPS	23
34	HAMBURGER	22

City of New York

The common open-ended question : "*What dishes do you like and eat often?*
(With a probe: "*Any other dishes you like and eat often?*").

Example 3. Open questions in sample surveys

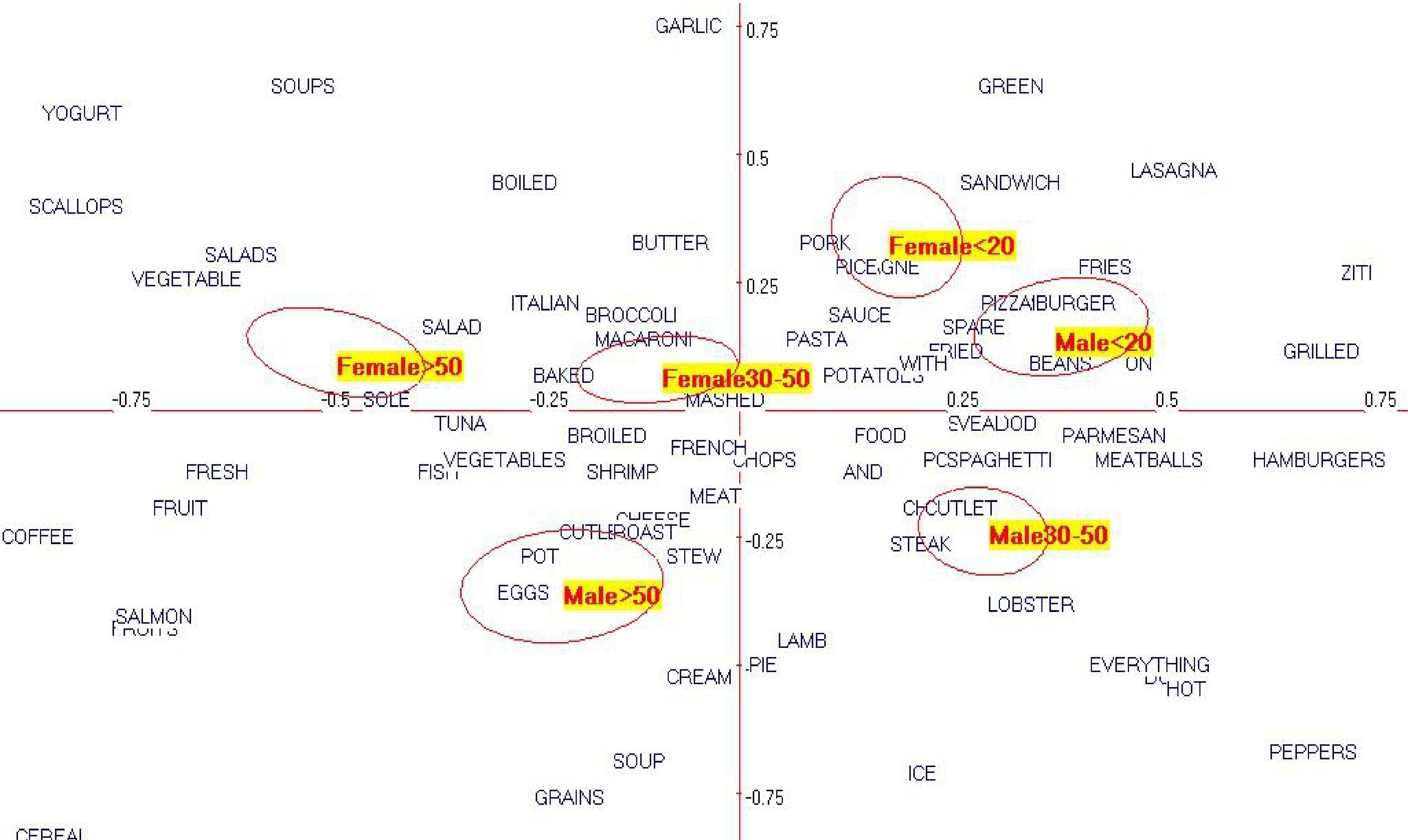
International survey (Tokyo Gas Company). A survey in three cities (Tokyo, New York, Paris) about dietary habits. Open question: "*What dishes do you like and eat often?*



New York: First principal plane. Table crossing words and age x gender categories

Example 3. Open questions in sample surveys

International survey (continuation). Question: "*What dishes do you like and eat often?*



New York: First principal plane. Example of confidence areas for categories (Bootstrap)

Example 3. Open questions in sample surveys

***Japanese sample :** 1008 respondents, 6900 occurrences; 880 word-forms*

Example of 3 Tokyo responses (in Latin characters)

— NIMONO/EIYO NO BARANSU GA TORE, MITAME NI UTSUKUSHII KOTO.

— YASAI SUPU / ESA DE NAKU SHOKUJI DE ARU KOTO, KAZOKU SOROTTE SHOKUJI O TORU, ANKA DE ARU KOTO.

— NIZAKANA, SARADA, NABERYORI, CHUKARYORI / ANZEN NA ZAIRYO O TSUKAI BARANSU NO YOI OISHIIMONO O KAZOKU YUJIN TO TANOSHIKU.

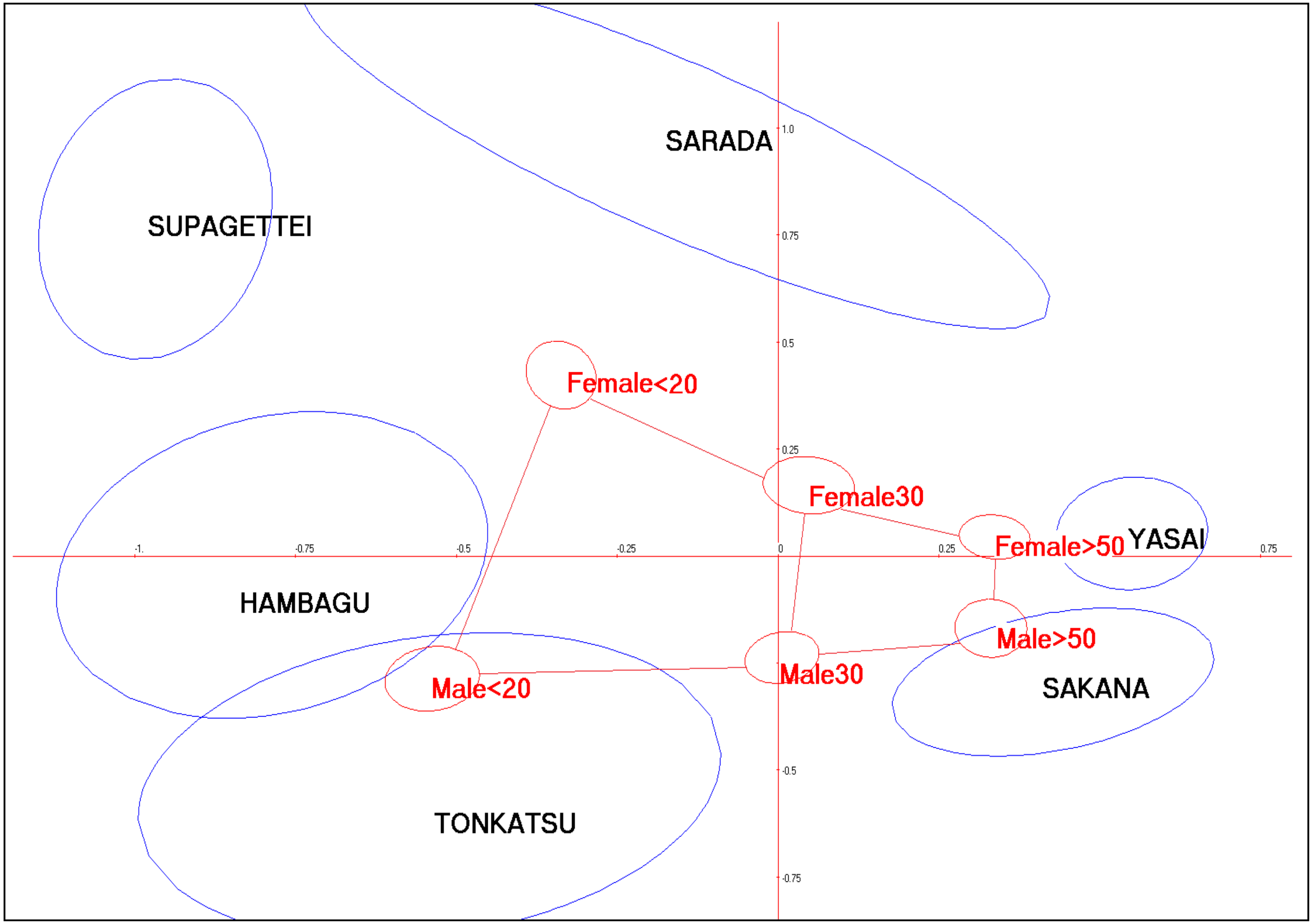
[- pot roast/balanced nutritionally, nice to look at.]

[- vegetable soup/ no food prepared sloppily, inexpensive things.]

[- cooked fish, salad, broth, Chinese food/ It is pleasant to eat with the family and with friends, good balanced meals with a lot of natural ingredients.]

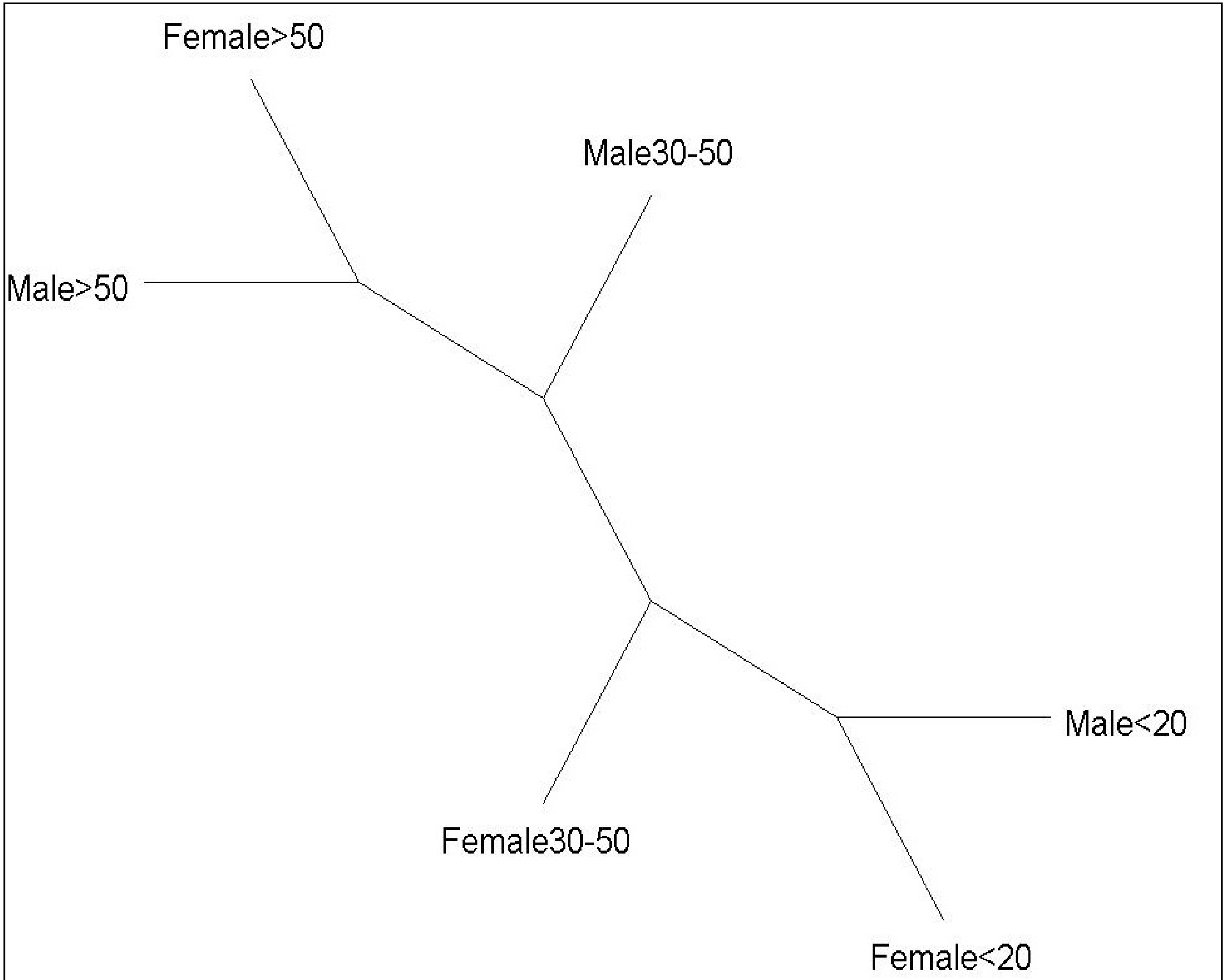
Example 3. Open questions in sample surveys

Location of 6 categories of respondents and some words (romanised characters)
Bootstrap confidence ellipses for both categories and words..



Example 3. Open questions in sample surveys

Japanese survey: Additive tree for 6 categories of respondents



Example 4. Synonyms of French verbs

Experience described in the book "La Sémiométrie" (Lebart, Piron, Steiner, Dunod, 2003) (or: *"The Semiometric Challenge",* freely downloadable from www.dtmvic.com) describing all the usual French verbs (the 829 most frequent verbs appearing in the classic grammar manual "Bescherelle") by all of their synonyms.

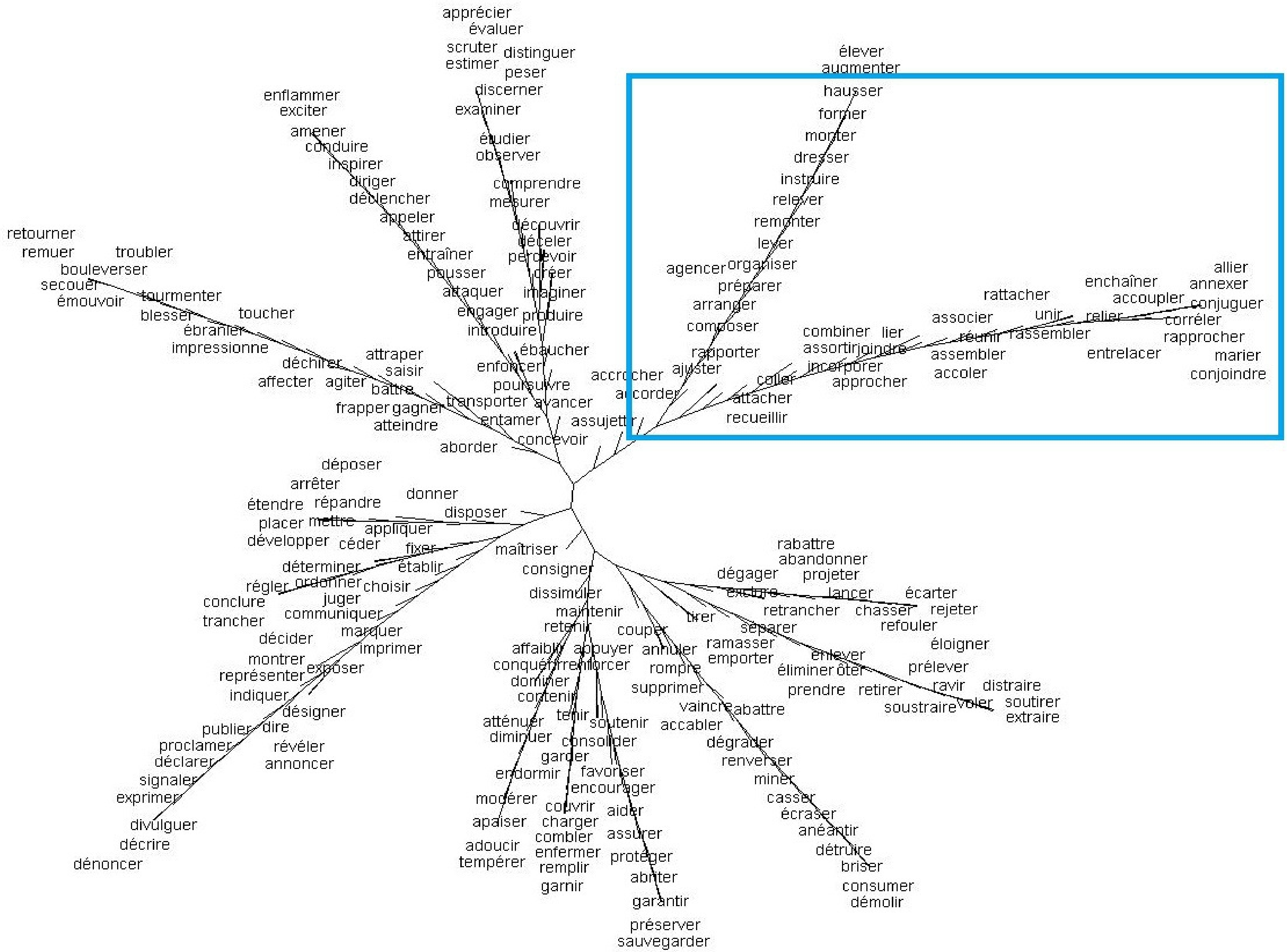
The "corpus" formed by verbs and their synonyms includes 17,446 occurrences of 3,839 distinct verbs. This number is greater than the 829 original verbs since less commonly used verbs can appear among the synonyms. We will treat below the 229 verbs having at least 20 synonyms

See also:

Ploux S. et Victorri B., Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement automatique des langues*, 39, n°1, 1998, pp.161-182.

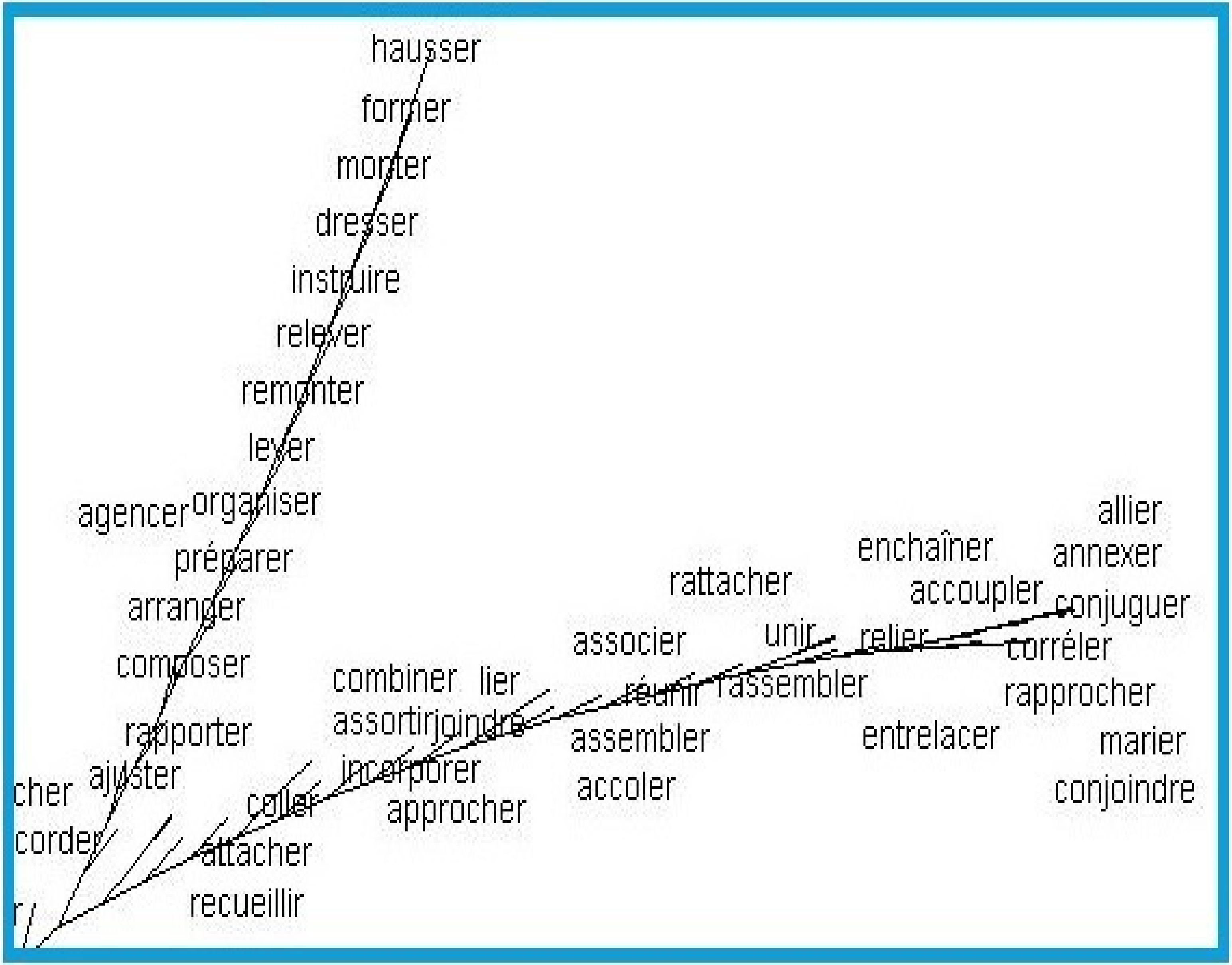
Gaume B., Venant F., Victorri B. Hierarchy in lexical organization of natural language, in D. Pumain (éd.), *Hierarchy in natural and social sciences*, Methodos series, vol 3, Springer, 2006, p. 121-142.

Example 4. Visualization of synonyms of 829 French verbs (additive tree)



Example 4. Visualization of synonyms of 829 French verbs

Figure . Zoom (blue windows of previous figure).



Example 4. Visualization of synonyms of 829 french verbs

The conclusions of the CA carried out on the same corpus were rather disappointing:

In fact, the geometric analysis of the multidimensional cloud of verb points shows that this cloud is almost spherical (neighboring eigenvalues).

This quasi-sphere has “lumps” on the periphery which are clusters of semantically neighboring verbs.

These “lumps” create the main axes according to their size, which also depends on the minimum frequency threshold chosen at the start.

This structure, painfully described by CA, is easily detected by additive trees.

Note that semantic similarity is not a transitive relationship

Example of semantic chains:

(1) *calm–wisdom–discretion–wariness–fear–panic*,

(2) *fact–feature –aspect–appearance–illusion* .

Example 4. Visualization of synonyms of 829 french verbs

Sattath and Tversky (1977), two of the founding fathers of **Additive Trees**.

" It is interesting to note that tree and spatial models are opposing in the sense that very simple configurations of one model are incompatible with the other model.
For example, a square grid in the plane cannot be adequately described by an additive tree."

Part 2. Simultaneous Additive Trees

We propose in this contribution a procedure for the simultaneous representation of texts and words for additive trees which makes it possible to combine:

the **advantages of some principal axis methods**

(simultaneous planar representation of rows and columns of a table)

and the **advantages of clustering techniques**

(better approximation of the distances in **full** space).

More generally, this procedure applies to the simultaneous representation of columns and rows of any contingency table.

2.1 Additive trees (AT): the phylogenetic explosion

These trees were originally proposed by Buneman (1971), then studied by Sattah and Tverski (1977).

A tree can be drawn with the objects as nodes (vertices), such as the distance between two objects is the length of the path joining these two objects on the tree.

More flexible than the MSP (Minimum Spanning Tree) which depends on n-1 parameter, the AT implies 2n - 3 parameters.

Saitou and Nei (1987) proposed an algorithm called **Neighbor Joining** which allows to roughly reduce the search of the additive tree to a classical ascending classification procedure.

This heuristic had a huge impact on the rapidly expanding world of phylogenetic research. Saitou and Nei's article has been cited more than 50,000 times since its publication.
Theoretical justifications for the algorithm's efficiency were presented by Mihaescu et al. (2009).

Simultaneous Additive Trees

The concept of hierarchy at the base of the ascending classification was to approximate the initial distances by an ultrametric distance, which satisfies, in addition to the classical axioms of any distance, for every triplet (x, y, z), the inequality:

$$d(x, y) <= Max(d(x, z), d(y, z)).$$

Additive trees are less demanding, although it is not obvious a priori, by asking only, for every quadruplet (x, y, z, t), that the inequality be verified:

$$d(x, y) + d(z, t) <= Max(\{d(x, z) + d(y, t)\}, \{d(x, t) + d(y, z)\})$$

With such a distance, a tree can be drawn with the objects as end elements (or leaves), such as the distance between two objects is the length of the path joining these two objects on the tree.

We can therefore have an idea of the **real distances** between elements on a **planar** graphical display.

Stimulated by the works of Barthélémy and Guénoche (1988) and Luong (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.

On additive trees drawing options

For a fairly complete review of general graph visualizations, one can consult Di Battista et al. (1999), and more particularly on methods using force-directed drawing algorithms, the article by Kobourov (2013) which analyzes more than 60 publications corresponding to several dozen algorithms.

Originally, the graph drawing algorithm of **Tutte (1963)** is one of the first drawing methods based on algorithms of this type.

The methods proposed by Eades (1984) and the algorithm of Fruchterman and Reingold (1991) are both based on repulsive forces between all the nodes of the graph, but also attractive forces between the nodes which are adjacent (the edges are assimilated to springs, and it is about finding a balance between all the tensions, hence the name force-directed drawings).

Alternatively, the forces between vertices can be calculated based on concepts from graph theory.

The distances between vertices are then the lengths of the shortest paths that join them.

For additive trees, these distances are precisely an approximation of the original distances (chi-square distances calculated on the original lexical table in the framework of textual data).

The algorithm of Kamada and Kawai (1989) uses these “spring forces” proportional to these distances calculated on the graph.

This tracing algorithm is therefore the most compatible with the properties of additive trees, and therefore with basic lexical distances.

Experimentally, we also note the good compatibility of these representations with the main plans resulting from the CA of the original lexical table.

2.2 Simultaneous representation in CA (reminder), characteristic words

Correspondence Analysis can be directly presented as the search for the best possible simultaneous representation of the proximity between rows and columns of a contingency table.

We can in fact look for an axis (to begin with) a simultaneous positioning of texts and words so as to obtain a doubly barycentric relationship: words at the barycenter of the texts, and texts at the barycenter of the words (the weights being respectively the lexical profiles of rows and columns calculated from the basic lexical table).

Evidently, this double relationship is impossible, because taking the barycenter is a shrinking transformation: the words must be inside the interval covered by the texts and, simultaneously, the texts inside the interval covered by the words.

For the relationship to be possible, the previous barycenters must be dilated (using a coefficient $\beta > 1$).

The optimal solution corresponds to a **value of β closest to 1.**

Such value gives us the positions of words and texts on the first axis of the CA of the basic table, and $\beta = (1/\lambda)^{1/2}$, λ being the largest eigenvalue of the CA.

For the axis α, $\beta_\alpha = (1/\lambda\alpha)^{1/2}$

If $\mathbf{v}_\alpha$ are the coordinate of the words (or rows)

If $\mathbf{u}_\alpha$ are the coordinates of the texts (or columns)

$$\begin{cases} \mathbf{v}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \mathbf{F}\mathbf{D}_p^{-1}\mathbf{u}_\alpha \\ \mathbf{u}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \mathbf{F}'\mathbf{D}_n^{-1}\mathbf{v}_\alpha \end{cases}$$

F is the frequency table

F' its transposed

Dp and **D**n the diagonal matrices containing the marginal frequencies

Note the simplicity of this presentation of CA obtained directly from doubly barycentric relationships known as **"transition relationships".**

The optimal solution corresponds to a **value of β closest to 1.**

Such value gives us the positions of words and texts on the first axis of the CA of the basic table, and $\beta = (1/\lambda)^{1/2}$, λ being the largest eigenvalue of the CA.

For the axis α, $\beta_\alpha = (1/\lambda_\alpha)^{1/2}$

If $\mathbf{v}_\alpha$ are the coordinate of the words (or rows)

If $\mathbf{u}_\alpha$ are the coordinates of the texts (or columns)

$$\begin{cases} \mathbf{v}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \mathbf{F}\mathbf{D}_p^{-1}\mathbf{u}_\alpha \\ \mathbf{u}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \mathbf{F}'\mathbf{D}_n^{-1}\mathbf{v}_\alpha \end{cases}$$

$\mathbf{F}$ is the frequency table

$\mathbf{F'}$ its transposed

$\mathbf{D}_p$ and $\mathbf{D}_n$ the diagonal matrices containing the marginal frequencies

Note the simplicity of this presentation of CA obtained directly from doubly barycentric relationships known as **"transition relationships".**

2.3 Drawing simultaneous trees

We cannot therefore hope to find an optimal double simultaneous representation, but a simple simultaneous representation is sufficient (words as barycenters of the texts, which are the vertices of the additive tree).

We will enrich this simultaneous representation by also bringing into play the notion of characteristic words (or specificities).

The simultaneous representation procedure that we propose includes the following steps: (columns and rows play similar roles and can be interchanged).

Sequence of computation steps

1) **Preliminary correspondence analysis** (of the contingency lexical table).

2) **Choice of the dimension *nx*** of the space deemed significant (generally through bootstrap validation) (12 axes for example). The distances will be calculated from the first *nx* main axes of the CA. (**This focus on significant principal space allows a regularization of the initial distances, a procedure well known in discriminant analysis and in certain Deep Learning procedures).**

3) **Computation of the additive tree** (Neighbors-Joining method) on the matrix of distances between the coordinates of the columns (texts) on the first *nx* axes.

4) **Drawing of the tree** (Kamada-Kawai procedure).

5) **Barycentric positioning of the rows** (words - forms, lemmas) from the coordinates of the column points (texts) (vertices of the tree) deduced from the procedure (4) and the textual profile of the rows (words/tokens/lemmas)).

6) **Computation**, directly from the lexical table, for each text column, **of the characteristic rows/words** (fixed probabilistic threshold) from the test-values (for the test-values, see for example: Lebart et al.; 1998, 2019).

7) **Drawing of new edges** (color and thickness different from those of the edges of the additive tree) joining each column point (text) on the graph to its characteristic lines (words).

These seven steps are in fact valid for all contingency tables.

In the case of textual data, a step “0” must be added to calculate the lexical table from the texts.

In the case of texts consisting of songs or poems, it is still necessary to add a preliminary “-1” step of converting the raw texts of the songs into “bags of words”.

The four examples of “augmented trees”

We will illustrate these “augmented trees” (additive trees with simultaneous representation of lines and columns) with applications to 4 corpora.

Example 5. “Inaugural address” corpus (State of the Union speeches)

Example 6. William Shakespeare : Sonnets

Example 7. The poet / singer Georges Brassens (194 songs)

Example 8. The poet / singer Leonard Cohen (80 songs)

The python code for the complete chain of the seven processing steps from raw texts to simultaneous visualizations of additive trees will be free and available.

Example 5. “Inaugural address” corpus

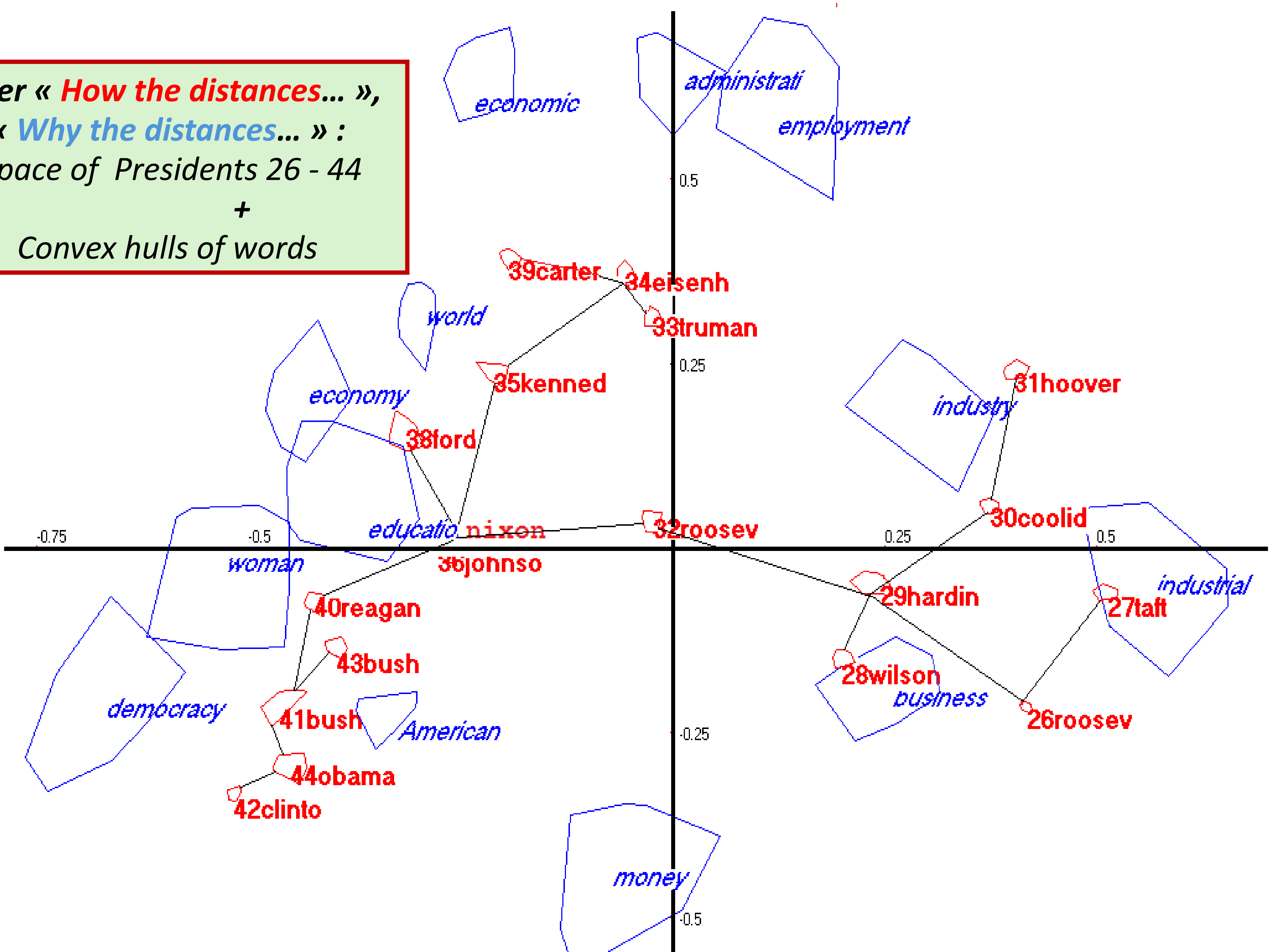
American Presidents SOTU speeches

State of the Union speeches of the 18 American presidents, excerpt from the “Inaugural address” corpus (that can be extracted from the nltk.book corpuses: see e.g. Bird et al. 2009)
[see also the website: http://www.usa-presidents.info/union/ that contains all the texts back from the speeches of George Washington in 1790].

As a check, the corpus was also lemmatized using the software TreeTagger (Schmid, 1994), with elimination of function words and prepositions.

Example 5. “Inaugural address” corpus. Example of CA with bootstrap convex hulls

After « How the distances… »,
« Why the distances… » :
Space of Presidents 26 - 44
+
Convex hulls of words

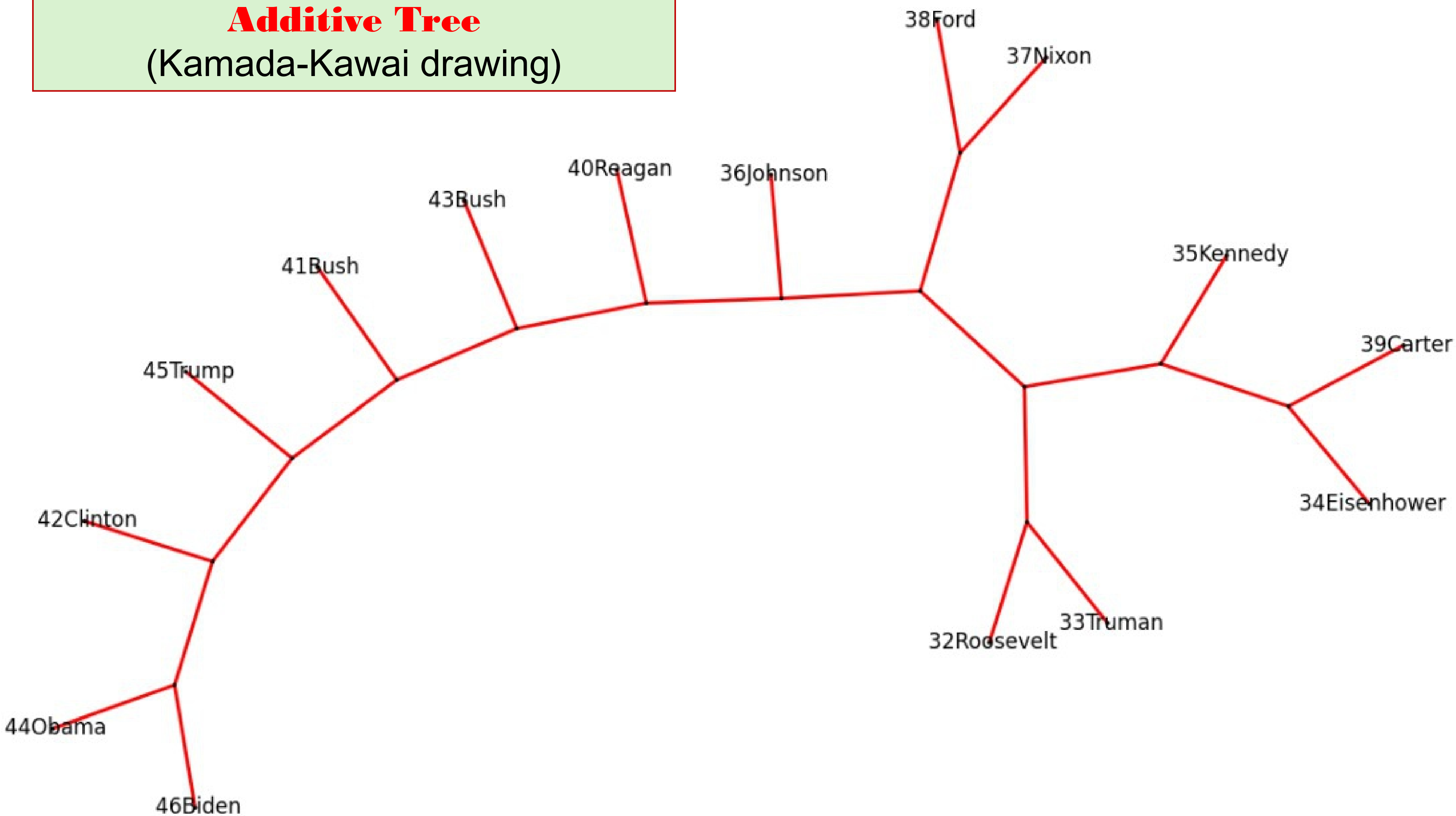


Example 5. “Inaugural address” corpus

State of the Union speeches 1942- 2024

Additive Tree

(Kamada-Kawai drawing)

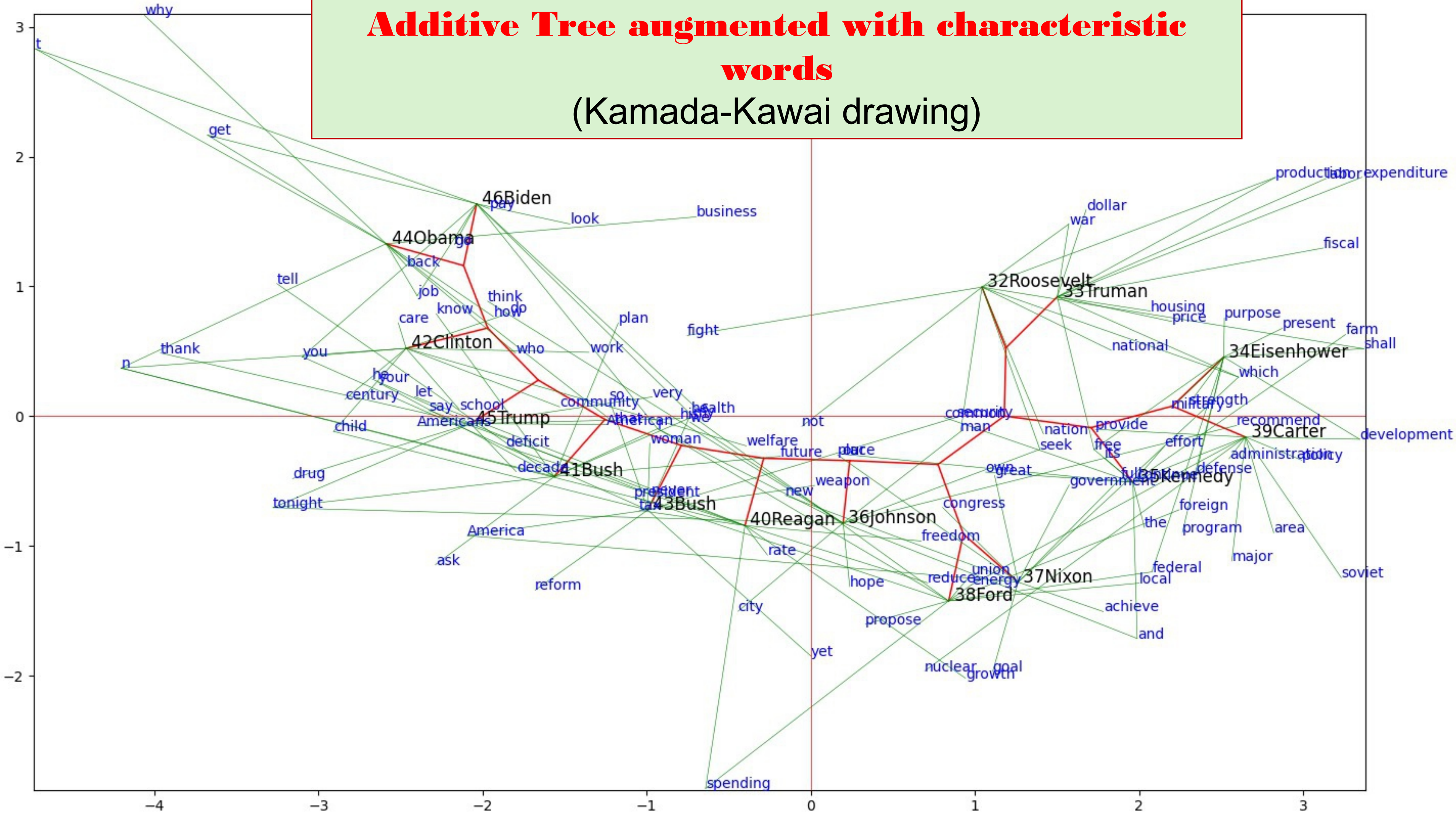


Example 5. “Inaugural address” corpus

State of the Union speeches 1942- 2024

Additive Tree augmented with characteristic words

(Kamada-Kawai drawing)



Example 6. William Shakespeare : Sonnets

The corpus of **Shakespeare's 154 Sonnets**, will serve as a reference body to present the Simultaneous Additive Trees (SAT). They are well known, translated into almost every language, deeply studied and commented.

Theme, Topic, Subject, Motif…

The definition of topics in Text Mining is pragmatic and can also cover the concepts of theme and motive (in the literary sense). Usually, a topic is the main subject, an objective explanation of the content of a text, while a theme represents a deeper underlying message. A motif is simply a recurrent idea used to reinforce the main theme. Schematically, the topics answer the questions: "What is the story about, who, what, how? "And the themes answer rather to:" Why was the story written? ". The topics in the literature are easier to identify than the themes.

Example 6. William Shakespeare : Sonnets

•The 154 sonnets of William Shakespeare deal with themes such as love, friendship, the effects of time, beauty, betrayal, lust, death. They are well known, translated into almost every language, deeply studied and commented

Three contiguous series of sonnets are generally recognized as corresponding to three dominant themes:

→ **Sonnets 1 to 17: (Procreation)**. These sonnets celebrate the beauty of a young man who is pressed by the poet to marry to perpetuate this beauty.

→ **Sonnets 18 to 126: (Young Man)**. This longest sequence concerns a young man (not identified), the destructive effect of time, the strength of love, friendship and poetry.

→ **Sonnets 127 to 154: (Dark Lady)**. These sonnets are mostly addressed to a dark haired woman. They are not devoid of irony or cynicism

Example 6. William Shakespeare : Sonnets

Eight themes inspired by expert comments

The themes Young Man and Dark Lady could themselves contain five sub-themes. The first theme (Procreation) remains as it is.

The new themes Young Man and Dark Lady include only the sonnets that are not assigned to the five new categories below (Absence, Storm, Rivalry, Death, Eternal poetry).

Table 1. Series of 8 themes / topics *a priori* followed by the sonnets numbers

Procreation	1 - 17
YoungMan	20-25, 33-38, 40-42, 46, 47, 49, 53-55, 59-60,62-70, 75-77, 88-106, 108-112, 115-125,
DarkLady	127-136, 139, 140, 143-146, 153,154
Absence	26-32, 39, 43-45, 48, 50-52, 56-58, 61, 113-114
Storm	141,142,147-152
Rivalry	78-87
Death	71-74
Etern_poetry	18, 19, 81

The partition of sonnets given in Table 1 is inspired by the works of Alden (1913) and Paterson (2010) but not explicitly mentioned by these authors.

Example 6. William Shakespeare : Sonnets

Sonnet 135… will and Will

whoever hath her wish, thou hast thy Will,
and Will to boot, and Will in overplus;
more than enough am I that vex thee still,
to thy sweet will making addition thus.
wilt thou, whose will is large and spacious,
not once vouchsafe to hide my will in thine?
shall will in others seem right gracious,
and in my will no fair acceptance shine?
the sea all water, yet receives rain still
and in abundance addeth to his store;
so thou, being rich in Will, add to thy Will
one will of mine, to make thy large Will more.
let no unkind, no fair beseechers kill;
think all but one, and me in that one Will.

Example 6. William Shakespeare : Sonnets

Shakespeare Sonnets.
Problems entailed by a
blind lemmatization:

Semantical drift, slang

house inn / vagina

weapon sword / penis

foin thrust

close with fight / embrace sexually

fist punch / masturbate

come advance / orgasm

vice grip

undone ruined financially / sexually, in terms of reputation

going departure / sexual activity

infinitive i.e. infinite, huge

thing item / penis

score tavern bill, accounts / vagina

Example 6. William Shakespeare : Sonnets

Locations of 7 *a priori* topics in the main plane of the CA of the lexical table (154 sonnets x 173 words), [min frequency = 10]. Here, the topics are supplementary variables (projected *a posteriori*)

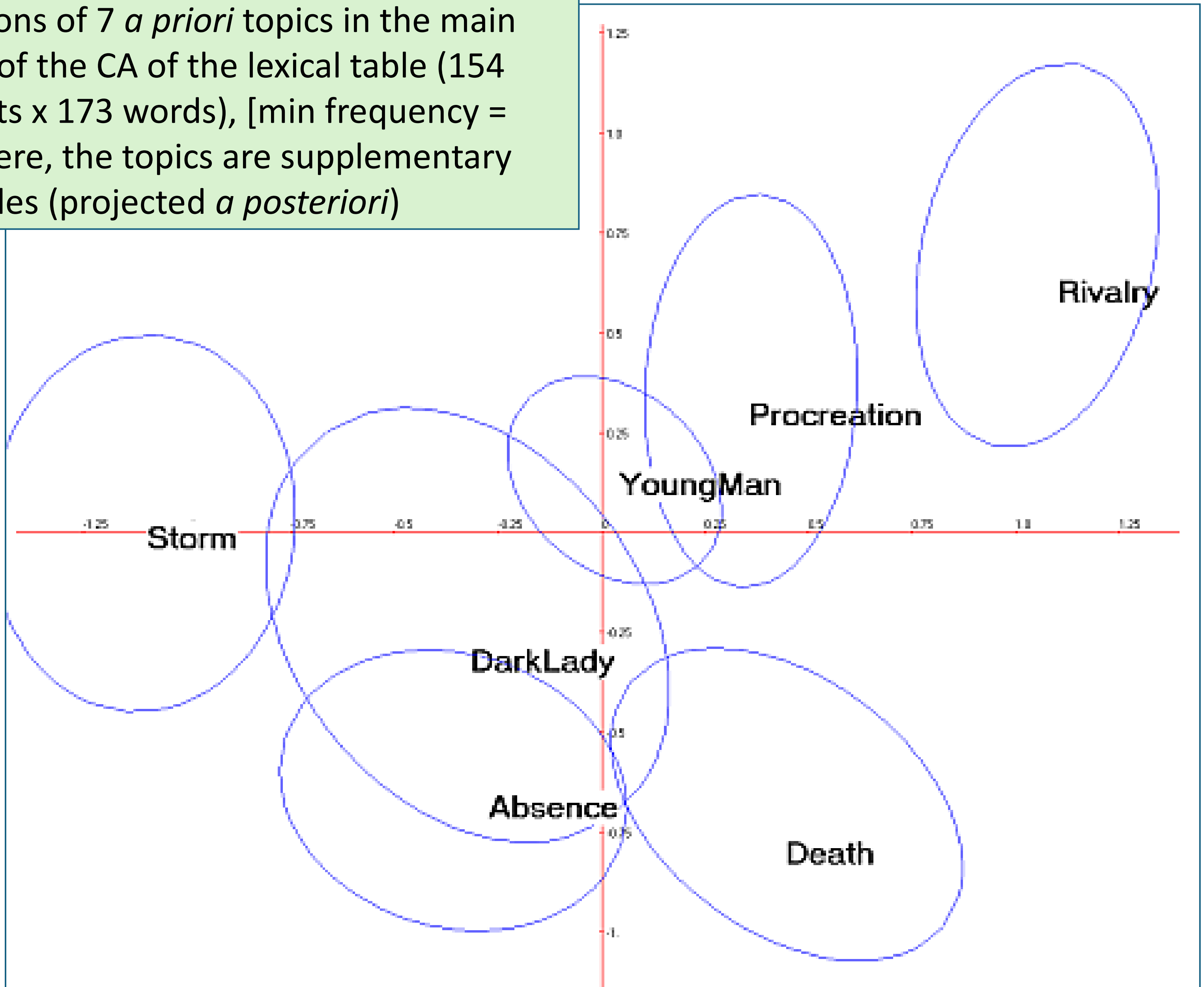


Table x. List of characteristics words for the 8 *a priori* topics

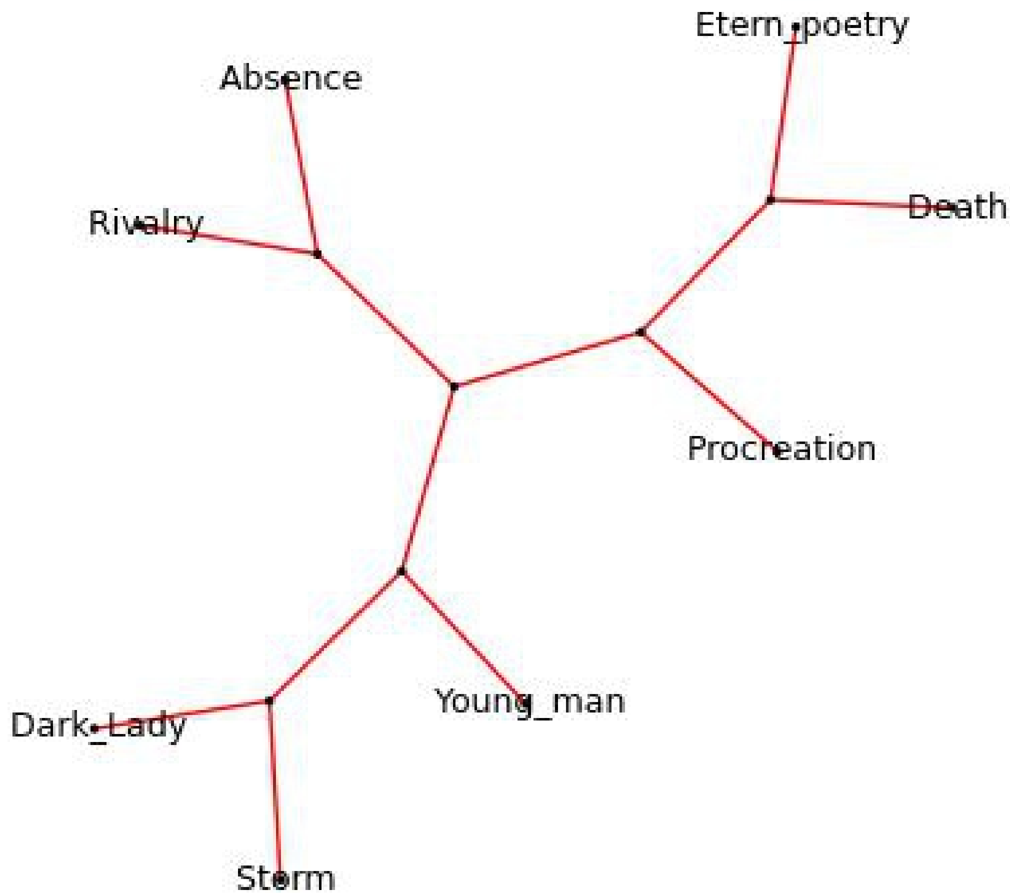
(Minimum frequency for words: 10, then, test-values > 1.7)

Procreation	beauty self world die age bear youth live time make
YoungMan	all never heart days time sun ever
DarkLady	black heart soul face one let well friend still
Absence	thought night day mind till being far woe think like
Storm	love eyes hate truth false see know best heart lies
Rivalry	praise worth making verse fair muse therefore use others
Death	world death would life
Etern_poetry	men long live world summer death

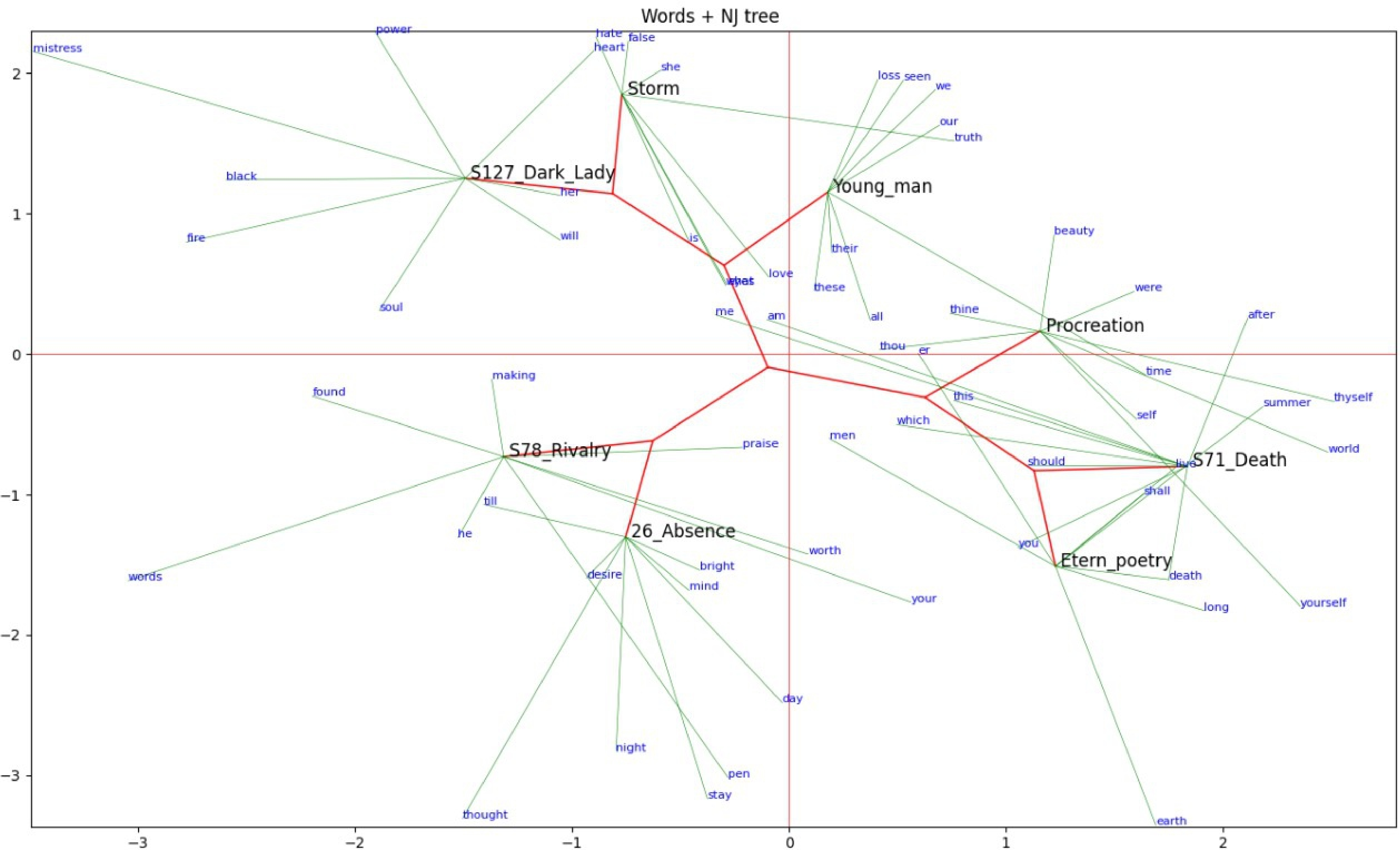
Example 6. William Shakespeare : Sonnets

Additive Tree

8 topics
(From 154 sonnets)

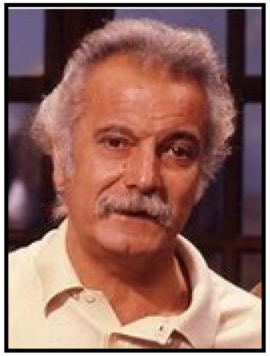


Example 6. William Shakespeare : Sonnets



Example 7. Georges Brassens (194 songs)

The corpus that we propose to study here is a more complex and elusive material: the collection of 194 songs written and sung by the French musician-poet Georges Brassens (1921-1981).



This author is special in that he brings together three almost contradictory features:

a) He was a nonconformist and has rubbed shoulders with anarchist movements,

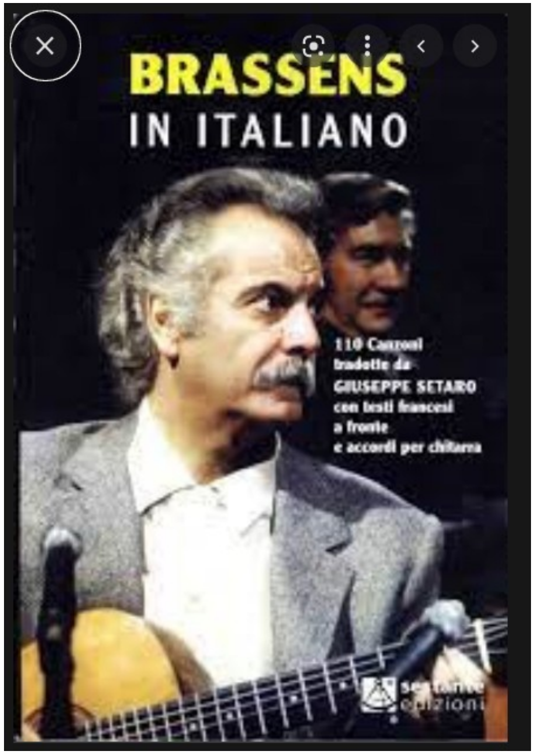
b) In 1967, he received the poetry prize from the very conservative *Académie Française*.

c) He was at the origin of the sale of 30 million records.

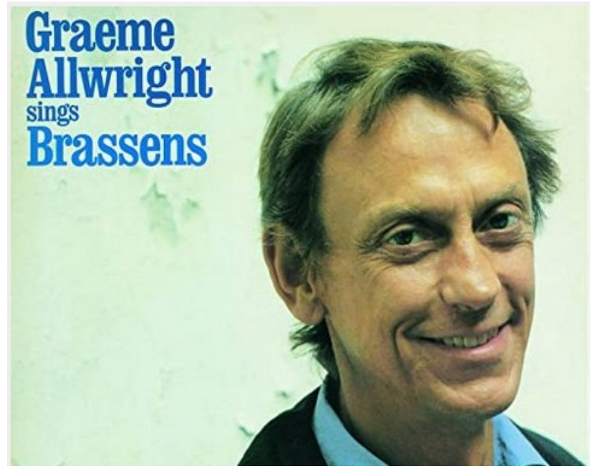
d)

Example 7. Georges Brassens (194 songs)

Italian



English



Japanese

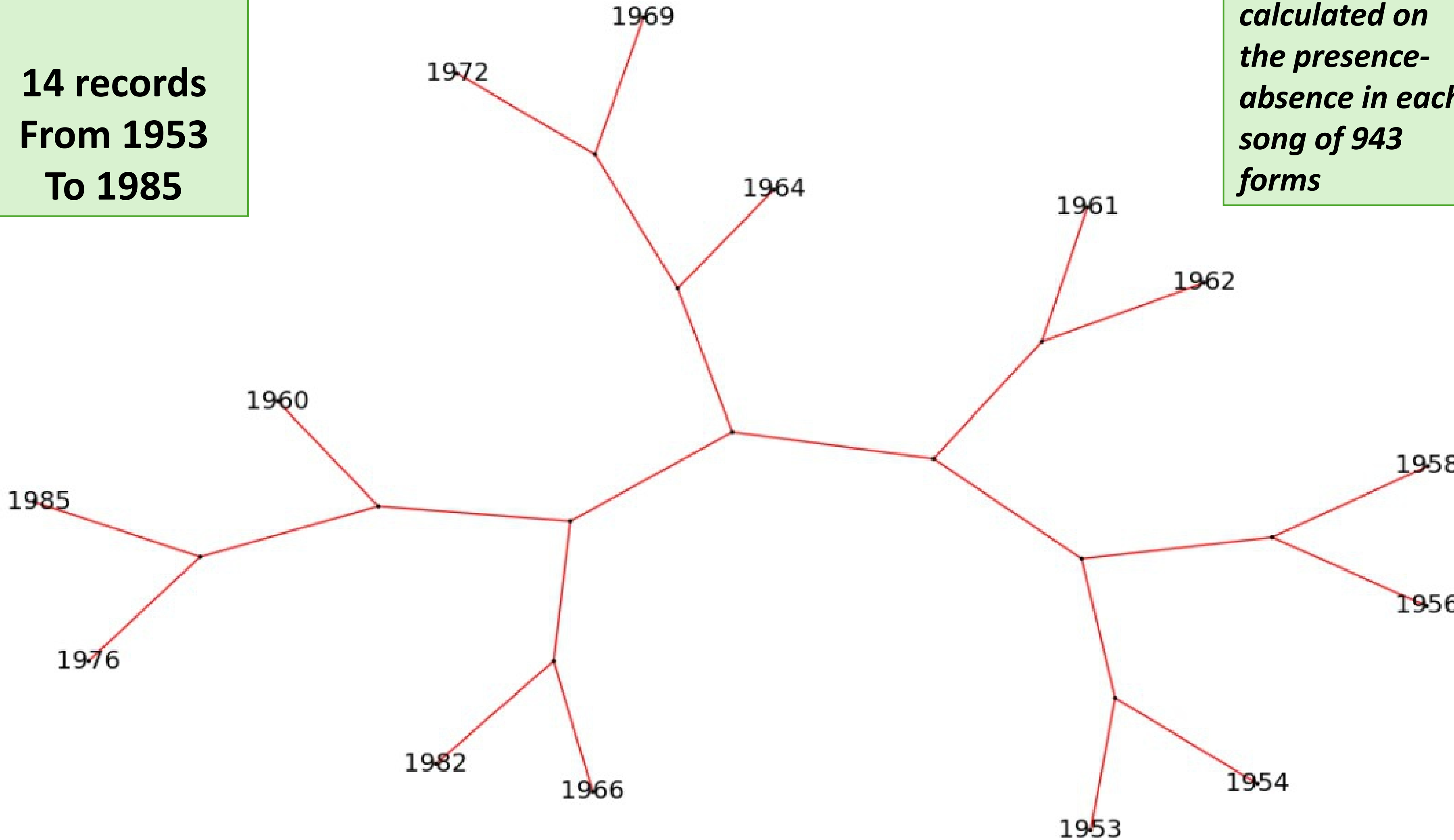
Spanish

Example 7. Georges Brassens (194 songs)

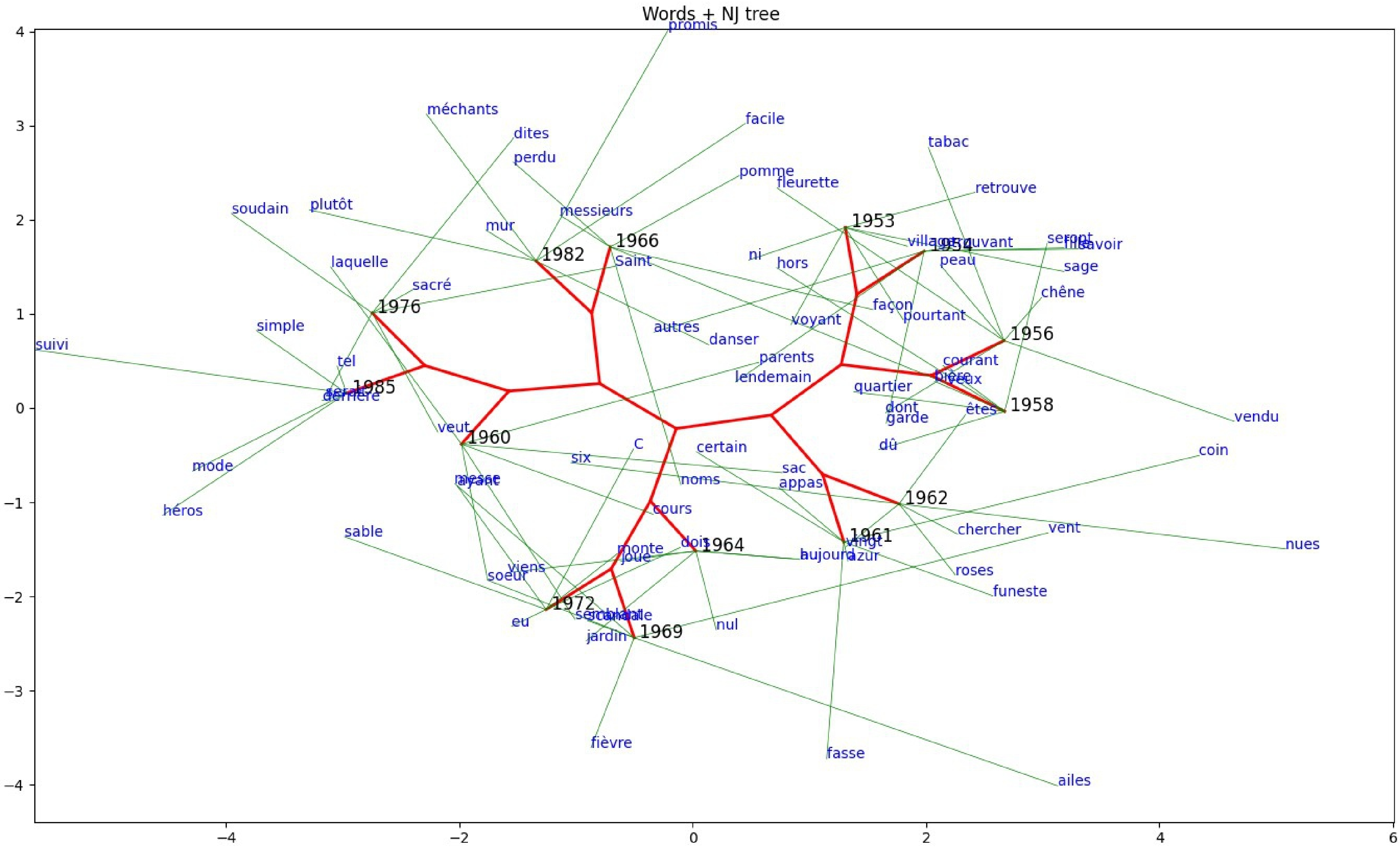
Additive Tree

14 records
From 1953
To 1985

Distances calculated on the presence-absence in each song of 943 forms



Example 7. Georges Brassens (194 songs)



Example 7. Georges Brassens (194 songs)

The statistical analysis of this type of text constitutes a methodological challenge.

Keep in mind the warning of Brunet (2004) during a statistical study of the poetic work of Arthur Rimbaud, a warning which applies also to the study of Brassens:

"… The use of … statistics does not go without a certain naivety which gives its faith to the printed words, in their first innocence. But with Rimbaud, the words are often loaded. They are decoys, figureheads, and the reality they designate and hide eludes the most learned interpretations.

When esotericism multiplies the traps, how to ensure the semantic constancy of the terms? ".

The studied works of Rimbaud included approximately 40,000 occurrences. The corpus of songs by Brassens used to exemplify the processing here has a comparable size: it contains approximately 52,000 occurrences.

Example 7. Georges Brassens (194 songs)

The poetic texts of Brassens are **particularly rich in stylistic figures (litotes, metaphors, anaphors, euphemisms, allegories, etc.)** which sow doubts about the use of the word (forms or lemmas) as a basic statistical unit.

This poet is an expert in the art of **diverting locutions** (he speaks of the "dark face of the honeymoon", of the "gospel according to Venus") (Lamy, 2004; Poulanges and Tilleu, 2001). It brings up to date **popular, outdated or slang expressions** (“the poor man's coffee” for: sexual act, etc.).

It often uses **historical, medieval and even obsolete** terms (Rochard, 2009).

In the case of songs that may include choruses or partial repetitions, the lexical frequencies no longer have the statistical significance given to them in the usual lexical tables. It is thus necessary to work with “bags of words” (presence- absence of words).

Example 8. Leonard Cohen (80 songs)

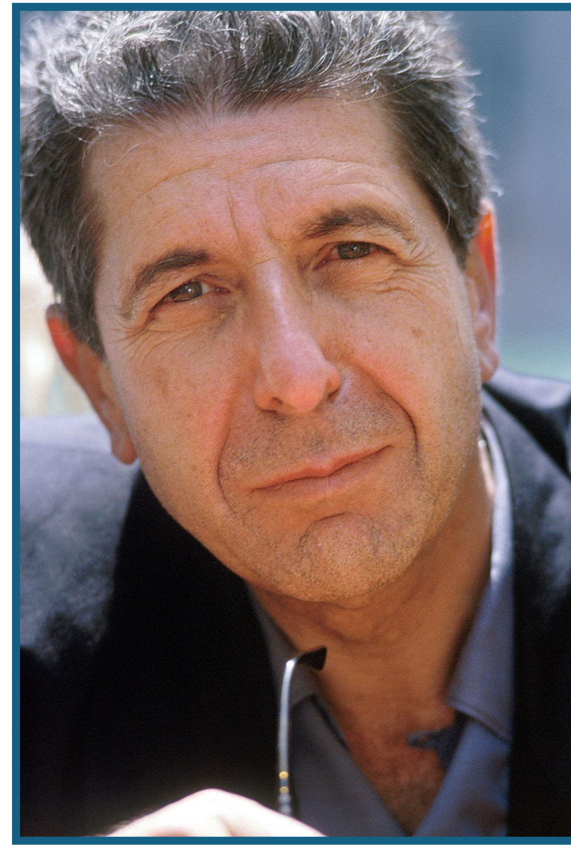
Leonard Cohen (1934 –2016) was a Canadian singer-songwriter, poet, and novelist.

Themes commonly explored throughout his work include faith and mortality, isolation and depression, betrayal and redemption, social and political conflict, and sexual and romantic love, desire, regret, and loss.

He was inducted into the Canadian Music Hall of Fame, the Canadian Songwriters Hall of Fame, and the Rock and Roll Hall of Fame.

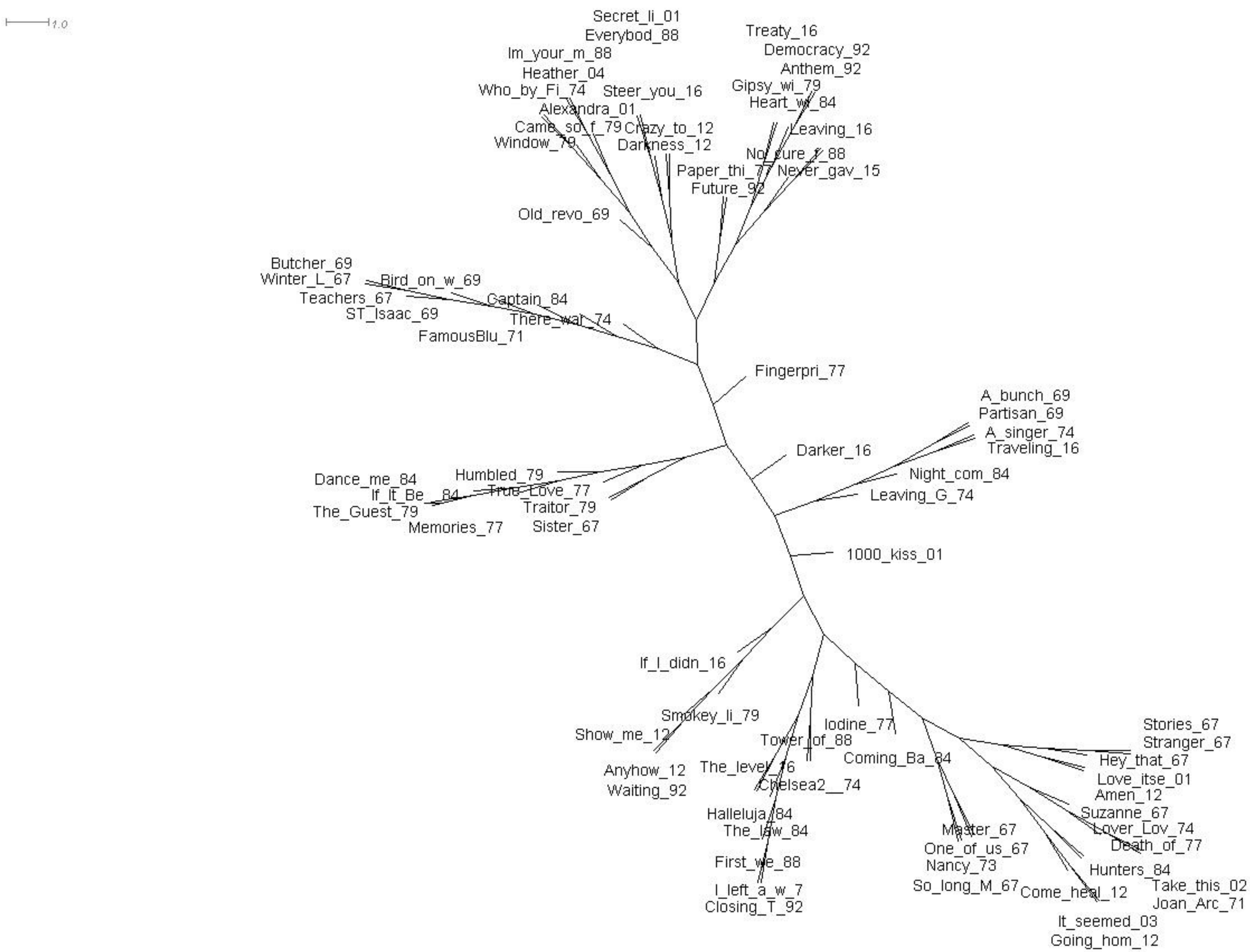
He was invested as a Companion of the Order of Canada, the nation's highest civilian honour.

In 2011, he received one of the Prince of Asturias Awards for literature and the ninth Glenn Gould Prize



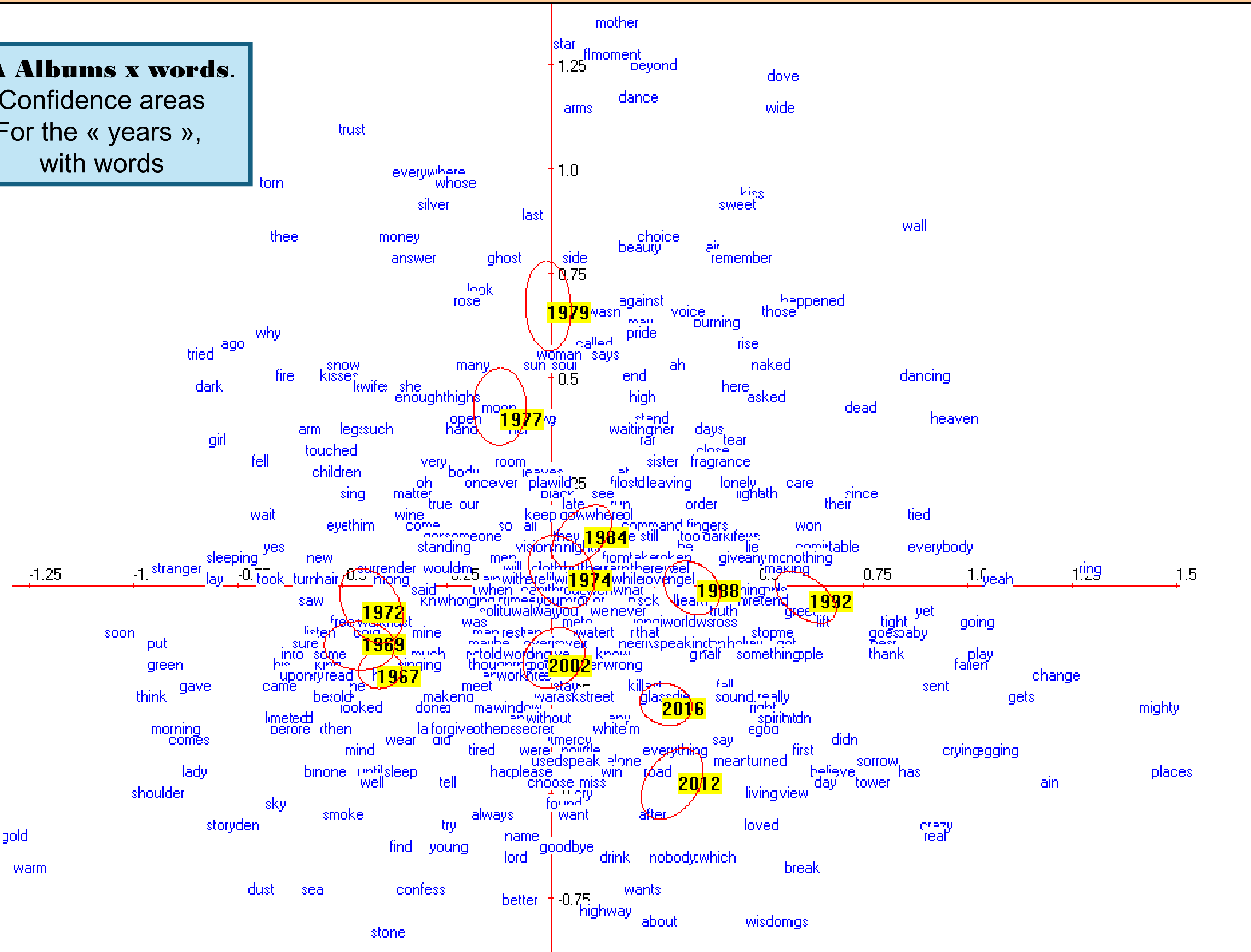
Example 8. Leonard Cohen (80 songs)

Example of a direct Additive tree of the 80 most popular songs



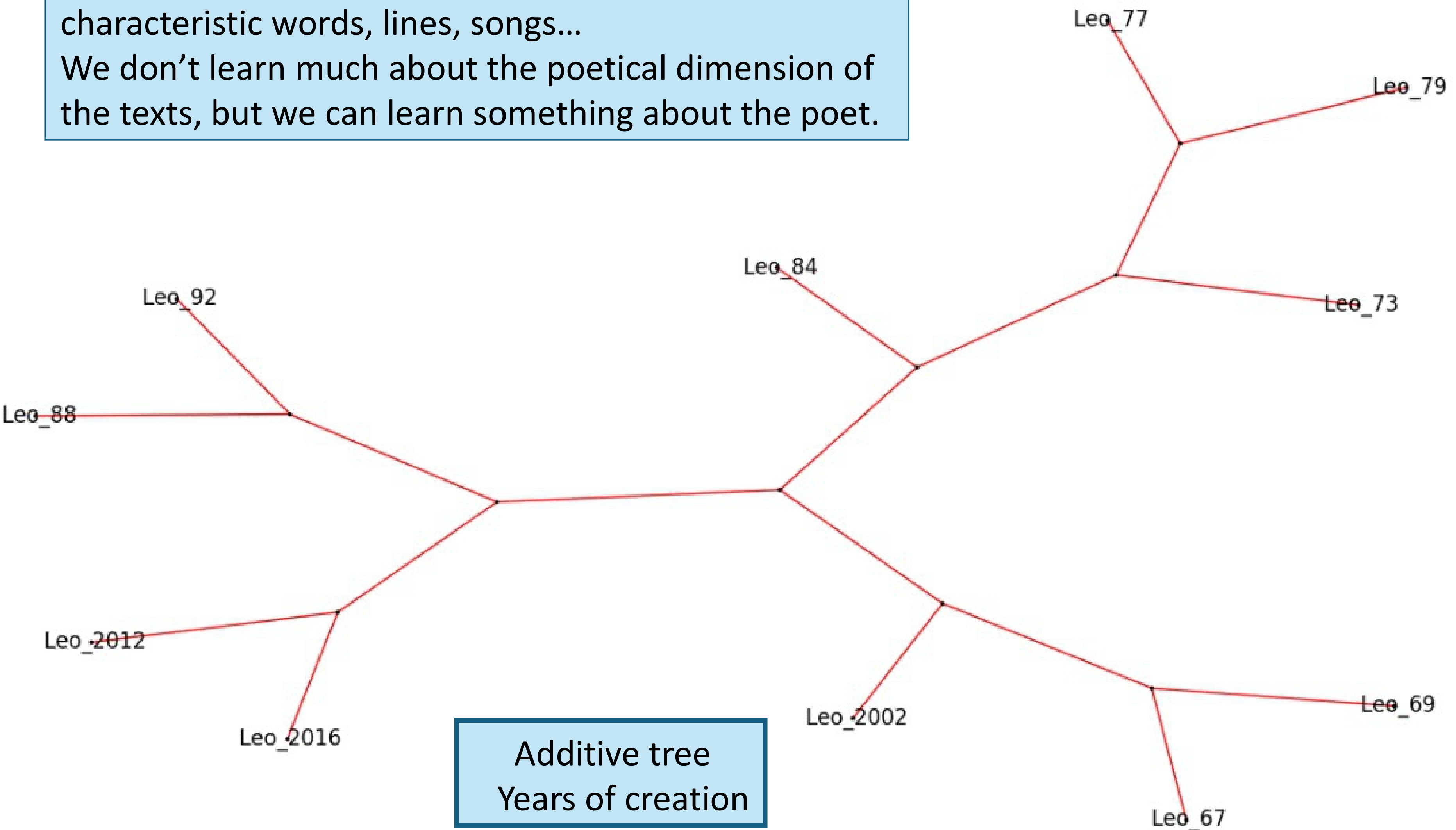
Example 8. Leonard Cohen (80 songs)

CA Albums x words.
Confidence areas
For the « years »,
with words



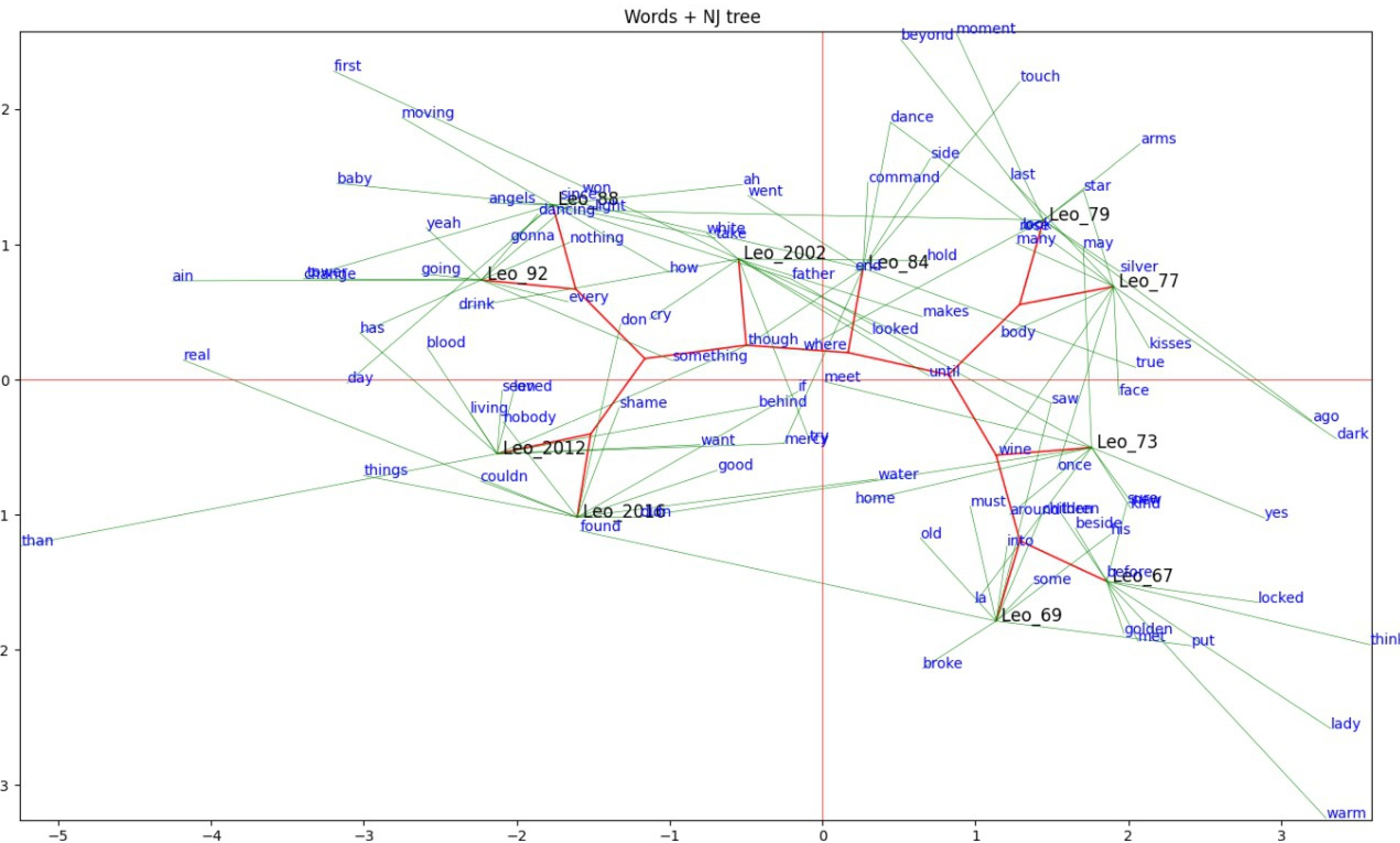
Example 8. Leonard Cohen (80 songs)

The undeniable trend can be documented by characteristic words, lines, songs…
We don't learn much about the poetical dimension of the texts, but we can learn something about the poet.



Additive tree
Years of creation

Example 8. Leonard Cohen (80 songs)



Conclurion. Current issues: Geometry, Parsimony, Validity, Usefulness

Reminder about Data Mining and KDD (Knowledge discovery from databases)

“ *Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* ” U.M.Fayyad, G.Piatetski-Shapiro

“ *I shall define Data Mining as the discovery of interesting, unexpected, or valuable structures in large data sets*” David. Hand

These two definitions use different words (*novel, unexpected, valid, useful, interesting, valuable, understandable, patterns, structures*):
all of them illustrate the difficulty to define precisely the exploratory approach

To assess the properties of an exploratory tool, we may check:

1 Its capacity to summarize and reconstitute some original data (Examples 1, 2)

2 Its ability to recognise patterns known beforehand (Examples 4,5,6)

3 Its ability to summarize, suggest, inspire (Example 3,7,8).

Conclusion. Again: Supervised or not...

Unsupervised learning had a catalytic effect in reviving interest in deep learning, but has since been overshadowed by the successes of purely supervised learning. Although we have not focused on it in this Review, 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.…

Le Cun, Bengio & Hinton, Deep Learning, Nature, 2015.
(this was perhaps the 46,539th citations…)

In the field of textual data analysis, the priority is not systematically "recognition" but discovery, description, comparison, understanding, observation of "statistical facts".

Such approach remains partially supervised in the sense that both the available external information and the discovered structures are used to enhance the exploration.

Data visualization methods certainly use algebraic or algorithmic methods similar to those of artificial intelligence. In some respect, CA is also a neuronal method (Lebart, 1997).

But a visualization is not a decision to be made, nor a task to be performed. It's almost the opposite. We don't ask questions to act, we submit data to understand and reflect.

The approach is unsupervised, a work phase which Deep Learning will increasingly need according to the predictions of the previous text by Le Cun et al. (2015).

As in correspondence analysis with its simultaneous representations, we have seen that the observable pattern of point-columns (texts, collections) tells us **how** these points are organized, and the presence of point-lines (words) tells us **why** they are organized in this way: texts are close because they often use the same words. And the words dress the skeleton of the additive tree.
But in addition to the practical difficulty of disseminating the real visualizations obtained (small formats, often: absence of color), there are the difficulties inherent in poetic texts and songs.

Conclusion. AI and poetry ?

The "argument from disability" makes the claim that "a machine can never do **X**." As examples of **X**, **Alan Turing** lists the following:
Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behavior as man, do something really new.

About poetry: Despite the limited number of graphical displays presented (necessarily small in size), one can guess that the textometric processing (multivariate description) of poetic texts brings a specific but **original point of view** on these texts, but above all about the authors, together with **new materials** for specialists.

From these first analyses, we were able to detect a general tendency, inextricably linked to age, career, personal development, perhaps to the growing notoriety of the poets and probably, (at least in the case of Brassens) to the increasing permissiveness during the period considered.

The use of word-forms may amplify, illustrate and nuance the results obtained from the lemmas. Each time, the use of characteristic elements reinforce the interpretations.

References (1)

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

Benzécri J.-P. (1973). *L'Analyse des Données*. Tome II : L'analyse des correspondances. Paris : Dunod.

Brunet É. (2004). Statistiques Rimbaldiennes, SI@T, *Les littératures de l'Europe unie*, Cesenatico, Italie, 88-113, hal-01362731.

Bryant D. (2005). On the uniqueness of the selection criterion in Neighbor-Joining. *Journal of Classification,* vol. (22), 1: 3-16.

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, Edinburgh: 387-395.

Di Battista, G. Eades, P., Tamassia R., et Tollis, I.G. (1999). Graph Drawing: Algorithms for the Visualization of Graphs, Englewood Cliffs : Prentice. Hall.

Eades P. (1984). A heuristic for graph drawing. *Congressus Numerantium*, 42:149–160.

Fruchterman T. et Reingold E. (1991). Graph drawing by force-directed placement. *Softw. – Pract. Exp*., 21(11):1129–1164.

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

Kamada T. et Kawai S. (1989). An algorithm for drawing general undirected graphs. *Inform. Process. Lett.*, 31:7–15.

Kobourov, S. G. (2013). Force-Directed Drawing Algorithms. in: *Handbook on Graph Drawing and Visualization.* Chapman and Hall/CRC.

References (2)

LeCun Y, Bengio Y, Hinton G. (2015). Deep learning. *Nature*. May 28;521(7553):436-44.

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., Baba Y. (eds). Berlin : Springer, 423-430.

Lebart L. (2000). Contiguity Analysis and Classification. In: *Data Analysis*, (Wolfgang Gaul, Otto Opitz, Martin Schader, eds), Berlin : Springer, 233-244.

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

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

Lebart L., Salem A., Berry L. (1998). *Exploring Textual Data*. Dordrecht, Boston : Kluwer Academic Publisher.

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

Mihaescu R., Levy D. et 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, *Les Cahiers de l'Analyse des Données*, 3, 187-198.

Rochard L. (2009). *Les mots de Brassens*, Paris : Edition du Cherche Midi.

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

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

Tutte, W. T.(1963). How to draw a graph. *Proc. London Math. Society,* 13(52):743–768.

More on: www.dtmvic.com

Thank You, Carlo!

Gracias

Grazie

Obrigado

Merci

Danke

Choukrane

75