JADT 2024

17ème Journées Internationales d'Analyse Statistique des Données Textuelles

Bruxelles, 25 – 27 Juillet 2024

Des outils pour décrire les tables lexicales,
(et certains corpus de poèmes et de chansons) :

Les arbres additifs simultanés

Ludovic Lebart

CNRS (R), *ludovic@lebart.org*

Des outils pour décrire les tables lexicales,
(et certains corpus de poèmes et de chansons)

Les arbres additifs simultanés

(ou : augmentés)

Partie I. Classification et Arbres Additifs (AA)

1.1 Arbres additifs : Principes et propriétés
1.2 Le tracé des arbres additifs
1.3 Les arbres additifs augmentés (et éléments illustratifs)

Partie 2. Applications à 3 corpus de chansons

2.1 Corpus Georges Brassens
2.2 Corpus Charles Aznavour
2.3 Corpus Jacques Brel
2.4 Corpus global et complement

Conclusion

Des outils pour décrire certains corpus de poèmes et de chansons **:**

Les arbres additifs simultanés: Résumé

L'analyse statistique des textes poétiques et des chansons est un défi méthodologique. Refrains, répétitions contraintes de versification mettent en question la signification statistique des fréquences lexicales.

On doit alors travailler sur des « **sacs de mots** » (présence ou absence de vocables) : une partie **infime** de l'aspect artistique des textes.

Avec ce codage, la dimensionnalité et la sphéricité du nuage ne permettent pas de bonnes visualisations dans des plans.

L'analyse des correspondances (AC) devient insuffisante, malgré sa propriété de représentation simultanée des mots et des textes**.** **Le calcul d'arbres additifs s'impose.** Nous proposons une nouvelle procédure de **représentation simultanée des textes et des mots** pour les arbres additifs.

Cette procédure permet de cumuler les avantages de l'AC et des classifications. On illustrera très schématiquement ce nouvel outil par des applications aux corpus de trois paroliers/chanteurs et poètes francophones Brassens, Aznavour et Brel.

Tools for describing some corpora of poetry and lyrics

Simultaneous additive trees: Abstract

The statistical analysis of poetic texts and songs is a methodological challenge. Choruses, repetitions or versification constraints question the statistical significance of lexical frequencies.

We must then work on "**bags of words**" (presence or absence of words) which evidently constitute a tiny part of the artistic aspect of these texts.
With this coding, the dimensionality and sphericity of the data do not allow good visualizations in a low-dimensional space: Correspondence Analysis (CA), becomes insufficient.

The computation of **additive trees** is essential. We propose a new procedure for the **simultaneous representation of texts and words** for additive trees. Such procedure allows us to combine the advantages of CA and clustering.

We give 3 examples of application of this new tool to corpora of lyrics of 3 francophone iconic singers and poets: **Brassens**, **Aznavour** and **Brel**.

Classification et arbres additifs

- Le calcul d'arbres additifs, proposé à l'origine par Buneman (1971) a été amélioré par Sattath et Tversky (1977).

- Les travaux de Barthélémy et Guénoche (1988) et de Luong (1988) ont favorisé l'utilisation de ces méthodes (sous le nom d'analyses arborées) dans le champ des analyses de texte.

- Les arbres additifs ont été rendus aisément calculables par Saitou et Nei (1987) et vont alors s'imposer comme synthèse entre les axes principaux et les techniques de clustering (classifications hiérarchiques, partitions type k-means, etc.).

- Bryant (2005), puis Huson et Bryant (2006) vont ensuite justifier la méthode et rendre accessible un logiciel correspondant. Des justifications théoriques de l'efficacité de l'algorithme ont été présentées par Mihaescu *et al.* (2009).

Le concept de hiérarchie à la base de la classification ascendante revenait à approximer les distances initiales par une distance *ultramétrique*, qui vérifie, en plus des axiomes classiques de toute distance, pour tout triplet (x, y, z), l'inégalité :

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

Les arbres additifs demandent, pour tout quadruplet (x, y, z, t), que soit vérifiée l'inégalité plus complexe et moins intuitive, mais moins exigeante :

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

Un arbre peut alors être dessiné avec les objets à classer comme éléments terminaux (ou « feuilles »). Plus souple que l'arbre de longueur minimale *(Minimum spanning tree)* qui dépend de *n-1* paramètres, l'arbre additif implique *2n – 3* paramètres.

Il faut donc trouver une approximation des distances initiales qui satisfasse ces conditions, ce que permettent les travaux précités de Saitou et Nei.

Le résultat important est que les distances entre nœuds (textes) du graphe (longueur du plus court chemin sur le graphe) sont des approximations des distances réelles dans tout l'espace, approximations souvent bien meilleures que celles fournies par un premier plan factoriel de AC.

Nous proposons donc dans cette contribution une procédure de représentation simultanée des textes et des mots pour les arbres additifs qui permet de cumuler les avantages des méthodes en axes principaux et des classifications.

Plus généralement, cette procédure s'applique à la représentation simultanée des colonnes et des lignes de toute table de contingence.

Il reste à tracer ces arbres.

Sur les options de traçage des arbres additifs

Pour une revue assez complète des visualisations de graphes généraux, on pourra consulter Di Battista *et al*. (1999), et plus particulièrement sur les méthodes utilisant les *force-directed drawings algorithms*, l'article de Kobourov (2013) qui analyse plus de 60 publications correspondant à plusieurs dizaines d'algorithmes.

A l'origine, l'algorithme de tracé de graphes de Tutte (1963) est l'une des premières méthodes de tracé fondée sur des algorithmes de ce type.

Puis les méthodes proposées par Eades (1984) et l'algorithme de Fruchterman et Reingold (1991) reposent toutes deux sur des forces répulsives entre tous les nœuds de l'arbre, mais aussi des forces attractives entre les nœuds qui sont adjacents (les arêtes sont assimilées à des ressorts, et il s'agit de trouver un équilibre entre toutes les tensions, d'où le nom de *force-directed drawings*).

1.2 Le tracé des arbres additifs

Alternativement, les forces entre les nœuds peuvent être calculées sur la base de concepts de la théorie des graphes.

Les distances entre nœuds sont alors les longueurs des plus courts chemins qui les joignent.

Pour les arbres additifs, ces distances sont justement une approximation des distances originales (distances du chi-2 calculées sur la table lexicale originale).

L'algorithme de Kamada et Kawai (1989) utilise ces « forces de ressort » proportionnelles à ces distances calculées sur le graphe.

Cet algorithme de tracé est donc le plus compatible avec les propriétés des arbres additifs, et donc avec les distances lexicales de base.

Expérimentalement, on constate d'ailleurs la bonne compatibilité de ces représentations avec les plans principaux issus de l'AC du tableau lexical original.

1.2 Le tracé des arbres additifs

Les capacités de l'analyse des correspondances à décrire des graphes non-orientés du type grille plane ou carte géographique à partir de leurs matrices associées ont été soulignées par Benzécri (1973, chap. 10) et Lebart *et al.* (1998).

On aurait pu penser utiliser de nouveau l'AC pour obtenir un tracé de l'arbre additif.

Mais les arbres (graphes connexes sans cycle) sont mal visualisés par cette méthode, et réciproquement, comme le notent Sattath et Tversky (1977)

"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."

Finalement, nous partirons donc des représentations fournies par la méthode de Kamada-Kawai la plus adaptée pour les arbres additifs, pour les enrichir par une représentation simultanée des lignes et colonnes de la table lexicale.

Représentation simultanée et mots caractéristiques

On peut présenter directement l'Analyse des Correspondances comme la recherche de la meilleure représentation simultanée possible des proximités entre lignes et colonnes d'une table de contingence (Lebart *et al*., 1984).

On peut en effet chercher sur un axe (pour commencer) un positionnement simultané des textes et des mots de façon à obtenir une relation doublement barycentrique : mots au barycentre des textes, et textes au barycentre des mots (les poids étant respectivement les profils lexicaux en ligne et en colonne calculés sur la table lexicale de base).

Cette double relation est impossible, car la prise de barycentre est contractante : les mots doivent être à l'intérieur de l'intervalle couvert par les textes et, simultanément, les textes à l'intérieur de l'intervalle couvert par les mots.

Pour que la relation soit possible, il faut dilater ces barycentres (coefficient b > 1). La solution optimale correspond à une valeur de b la plus proche de 1 qui nous donne les positionnements des mots et des textes sur le premier axe de l'analyse des correspondances. On notera la simplicité de cette présentation de l'AC obtenue directement à partir des relations doublement barycentriques connues sous le nom de « relations de transitions ».

La procédure de représentation simultanée que nous proposons :

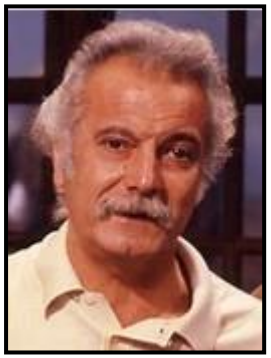
1) Analyse des correspondances préliminaire de la table lexicale.
2) Choix de la dimension *nx* de l'espace jugé significatif (en général par *bootstrap*) (12 axes par exemple). Les distances seront calculées à partir des *nx* premiers axes principaux de l'AC. **[**Régularisation des distances initiales, procédure bien connue en **analyse discriminante** et en ***Deep Learning***.]
3) Calcul de l'arbre additif (*Neighbors-Joining method*) sur la matrice des distances ainsi calculée.
4) Tracé de l'arbre (procédure de Kamada-Kawai).
5) Positionnement barycentrique des lignes (mots - formes, lemmes) à partir des coordonnées des nœuds de l'arbre (textes) (points-colonnes) et du profil textuel des lignes (mots/graphies/lemmes).
6) Calcul, à partir de la table lexicale, pour chaque texte, des lignes/mots caractéristiques (seuil probabiliste fixé) à partir des valeurs-test.
7) Tracés de nouvelles arêtes (couleur et épaisseur différentes de celles des arêtes de l'arbre additif) joignant sur le graphe chaque point-colonne (texte) à ses lignes (mots) caractéristiques.

Ces sept étapes sont en fait valables pour toutes tables de contingence. Dans le cas de données textuelles, il faut ajouter une étape « 0 » de calcul de la table lexicale à partir des textes.

Dans le cas des textes formés de chansons ou de poèmes, il faut encore ajouter une étape « -1 » préliminaire de conversion des textes bruts des chansons en « sacs de mots » (2 lignes de code en *Python*…)

On illustrera ces « arbres augmentés » (arbres additifs avec représentation simultanées des lignes et des colonnes) avec des applications aux corpus de trois musiciens, paroliers et poètes prédominants dans la chanson francophone du vingtième siècle : Georges Brassens, Charles Aznavour et Jacques Brel.

2.1 Corpus Georges Brassens (194 chansons)



Nous évoquons ici le recueil de 194 chansons chantées et enregistrées par le musicien–poète français Georges Brassens (1921–1981) regroupées en 14 recueils correspondant à autant d'albums (disques).

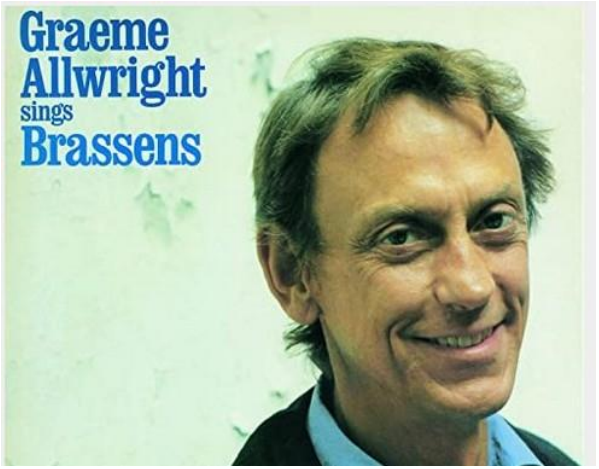
Cet auteur non-conformiste, qui a fréquenté les mouvements anarchistes, a cependant reçu en 1967 le prix de poésie de l'Académie Française. Traduit en plusieurs langues (anglais, italien, allemand, japonais…), il a été à l'origine de la vente de plusieurs dizaines de millions de disques.

Les textes poétiques de Brassens sont particulièrement riches en figures de style (litotes, métaphores, anaphores, euphémismes, allégories, …) qui posent des problèmes lors de l'utilisation du mot (graphie ou lemme) comme unité statistique de base (Rochard, 2009).

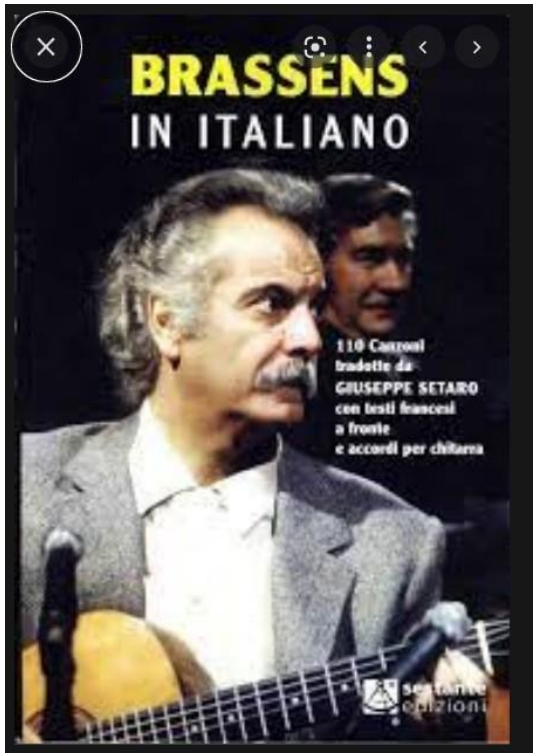
2.1 Corpus Georges Brassens (194 chansons)

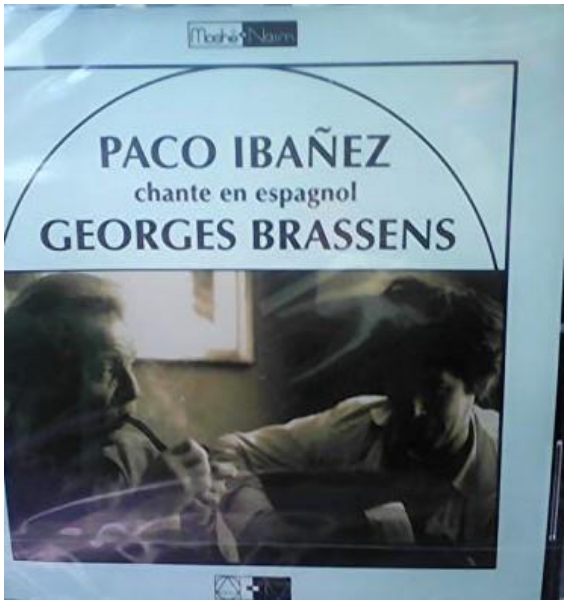
English

Italian



Spanish





Japanese

2.1 Corpus Georges Brassens (194 chansons)

Pour remédier à cette défaillance statistique des fréquences des formes graphiques (graphies) ou des lemmes, on va donc transformer chaque texte de chanson en vocabulaire non pondéré, autrement dit, chaque élément n'apparaitra qu'une seule fois à l'intérieur d'une chanson donnée (*words bag*).

Comme souvent en analyses textuelles, on aura en fait deux jeux de données provenant d'une part du fichier brut, d'autre part du fichier lemmatisé.

Le fichier lemmatisé a l'avantage de réduire la diversité des flexions et donc de permettre des seuils de fréquence minimale plus bas.

Le fichier des graphies garde la diversité originale des formes ce qui est fondamental dans le cas de textes poétiques.

On obtiendra donc à chaque étape deux points de vue différents et complémentaires.

2.1 Corpus Georges Brassens (194 chansons)

Pour la plupart des analyses de type AC portant sur l'ensemble des 194 chansons, ou sur les 170 chansons dont Brassens est le seul auteur, une dimension est dominante, qu'il s'agisse de lemmes ou de graphies : elle oppose les textes anciens (quatre ou six premiers albums) aux textes plus récents.

Les textes anciens, comme les poèmes externes, ont un vocabulaire que l'on peut qualifier de classique, voire galant ou précieux, pour forcer le trait.

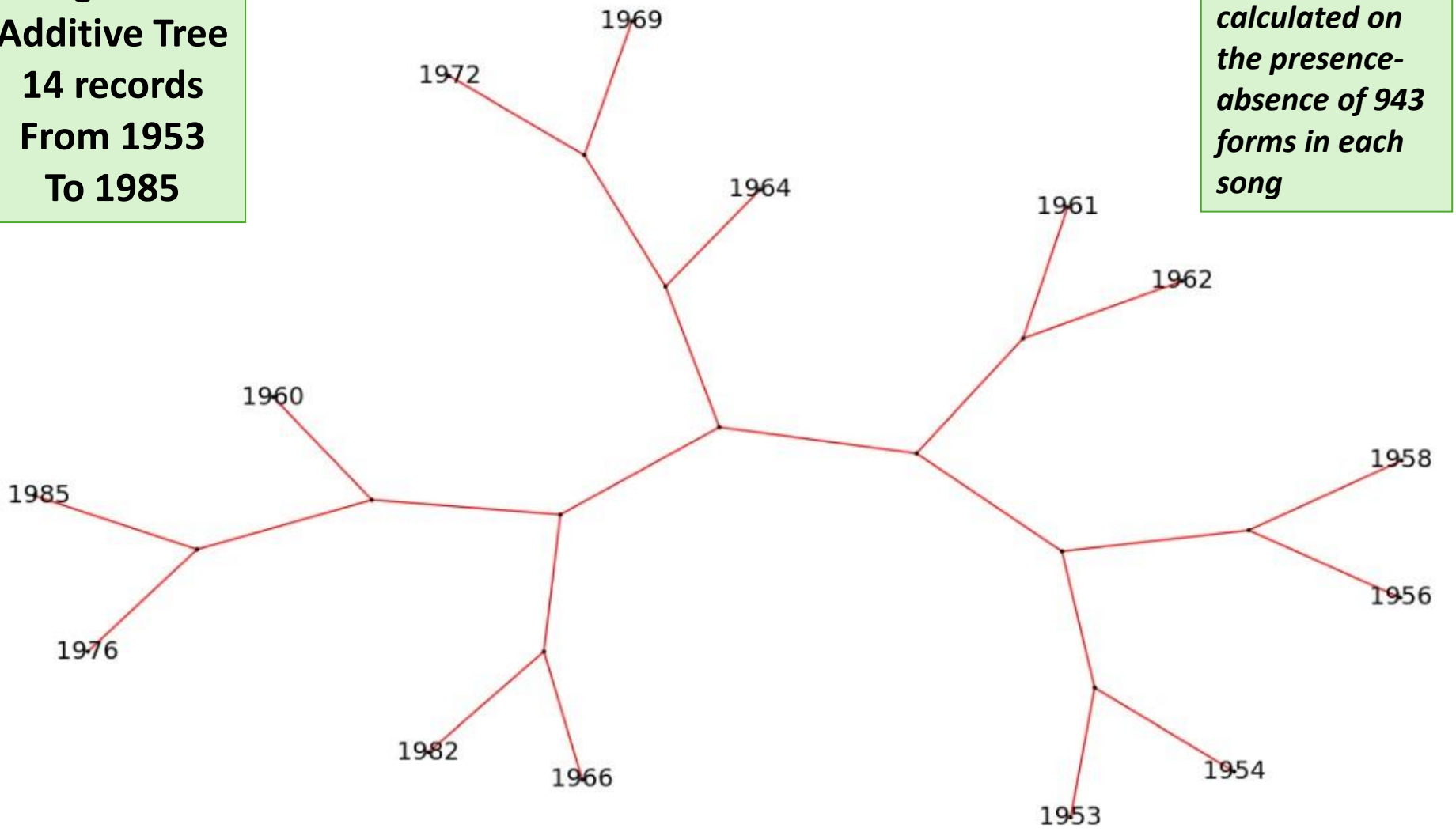
Les plus récents ont un vocabulaire plus cru, parfois argotique, provocateur, « salle de garde ».

La figure 1 nous donne le tracé de l'arbre additif « nu » avec ses 14 nœuds.

La table lexicale comporte 944 lignes (formes graphiques) et 14 colonnes (14 recueils des 170 chansons entièrement écrites et composées par Brassens).

2.1 Corpus Georges Brassens (194 chansons)

Figure 1
Additive Tree
14 records
From 1953
To 1985



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

2.1 Corpus Georges Brassens (194 chansons)

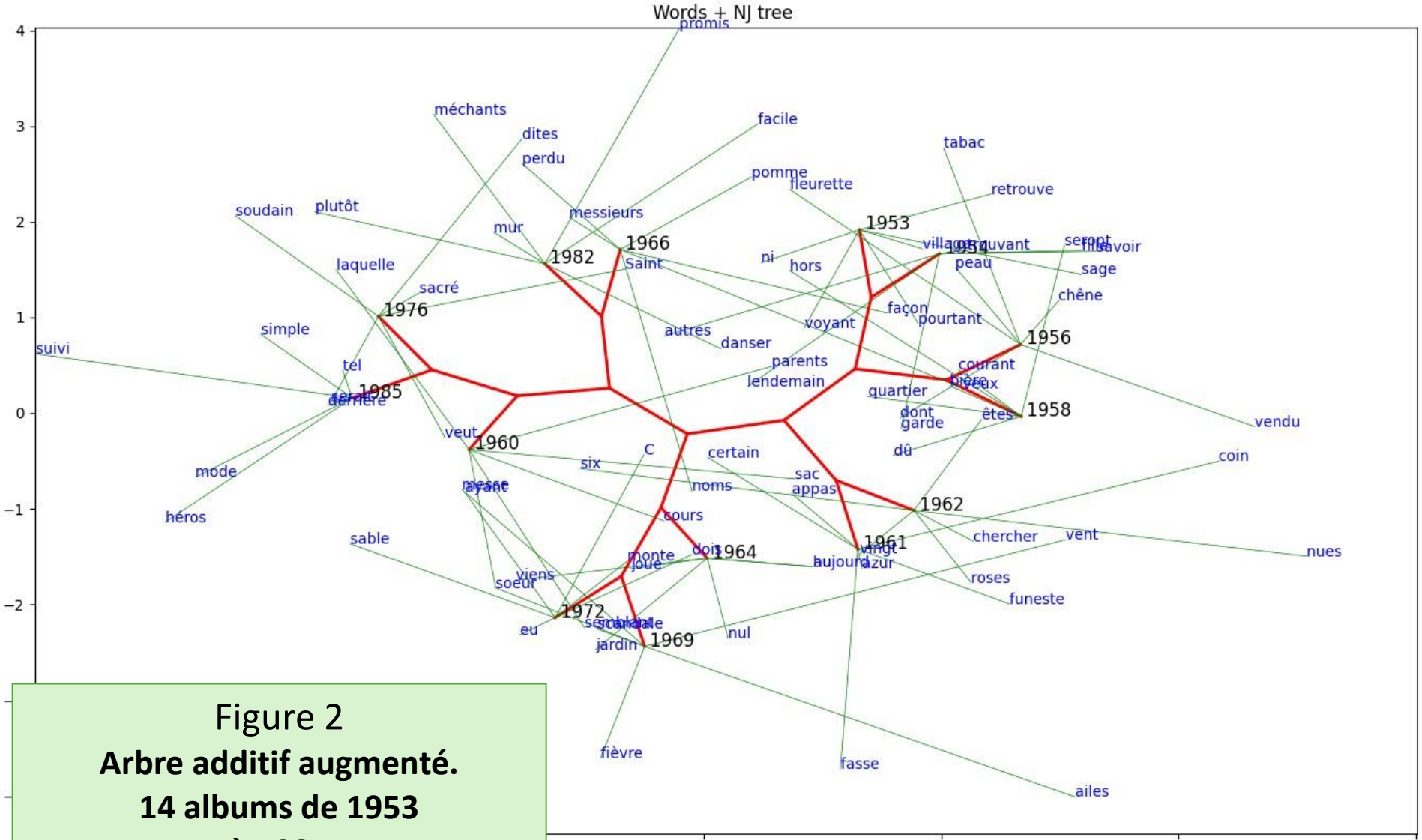


Figure 2

Arbre additif augmenté.
14 albums de 1953
à 1985

2.1 Corpus Georges Brassens (194 chansons)

La figure 2 qui reprend le tracé de l'arbre (à des changements d'orientation près) ne comporte qu'un tout petit extrait des graphies actives (70 au lieu de 944).

Pour chacun des 14 recueils analysés, on ne retient que les 5 graphies les plus caractéristiques au sens des valeurs-tests.

Ces critères, calculés directement à partir de la table lexicale de base, permettent d'obtenir les mots les plus caractéristiques de chaque recueil, avant toute analyse.

2.1 Corpus Georges Brassens (194 chansons)

La figure 3 présente à titre de comparaison avec la figure 2 le premier plan factoriel (AC) de la même table lexicale (944 x 14), avec le même petit sous-ensemble de graphies.

Ces graphies sont aussi jointes par un trait fin aux recueils (années) qu'elles caractérisent avant toute analyse.

Remarquons, par exemple que les premières années, 1954, 1956, 1958 sont presque superposées à droite dans le plan AC de la figure 3, alors qu'on peut lire clairement sur l'arbre additif de la figure 2 que 1953 est en fait le plus proche voisin de 1954, et que cette paire d'année se distingue de la paire (1956, 1958).[Rappelons que le plus court chemin sur l'arbre additif représente les vraies distances dans l'espace complet].

Même remarque pour le dernier recueil, assez hétéroclite et remanié par Jean Bertola en 1985 (quatre ans après le décès du poète) qui paraît proche de 1953 sur la figure 3 (AC, haut, droit) et largement opposé sur la figure 2 (extrême gauche et haut droit).

2.1 Corpus Georges Brassens (194 chansons)

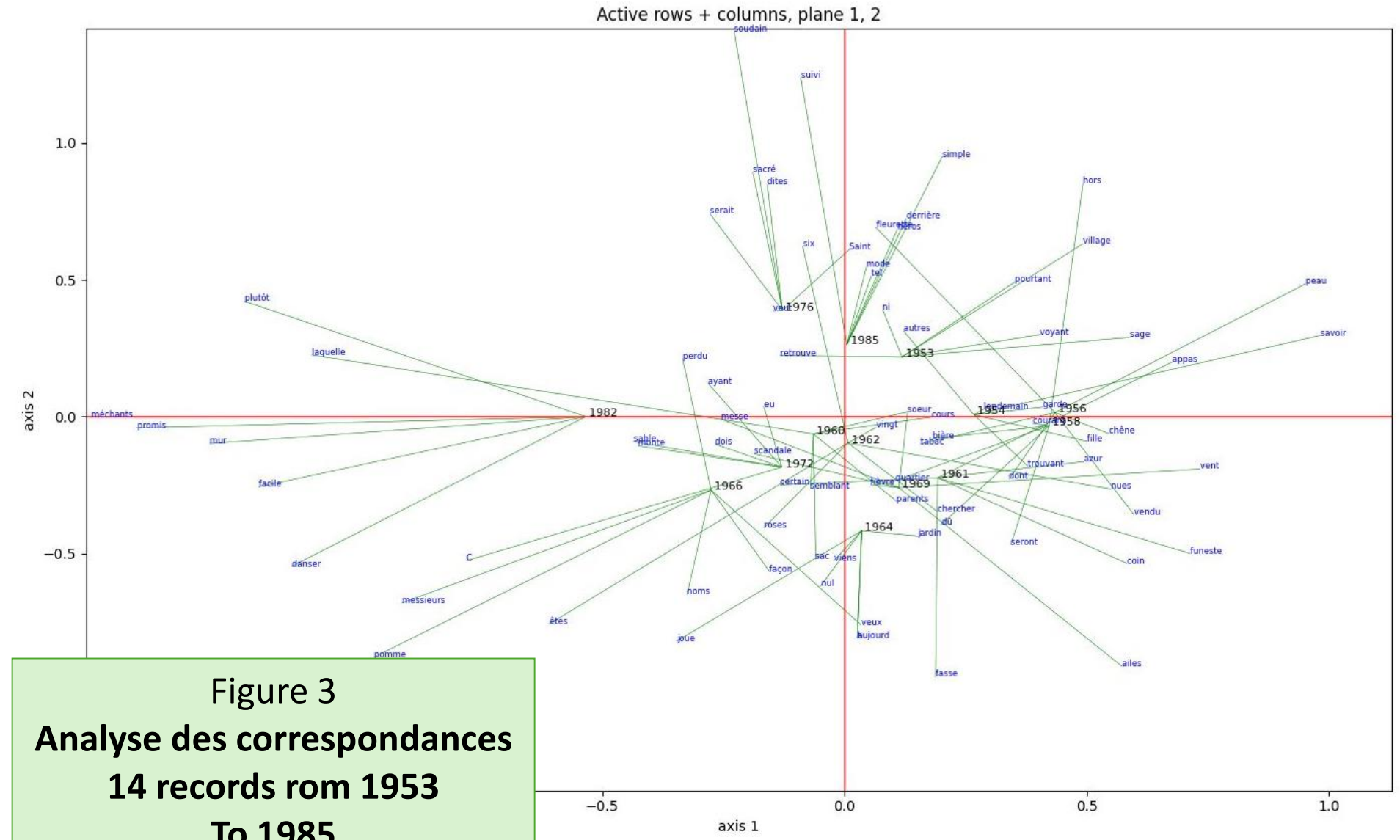


Figure 3

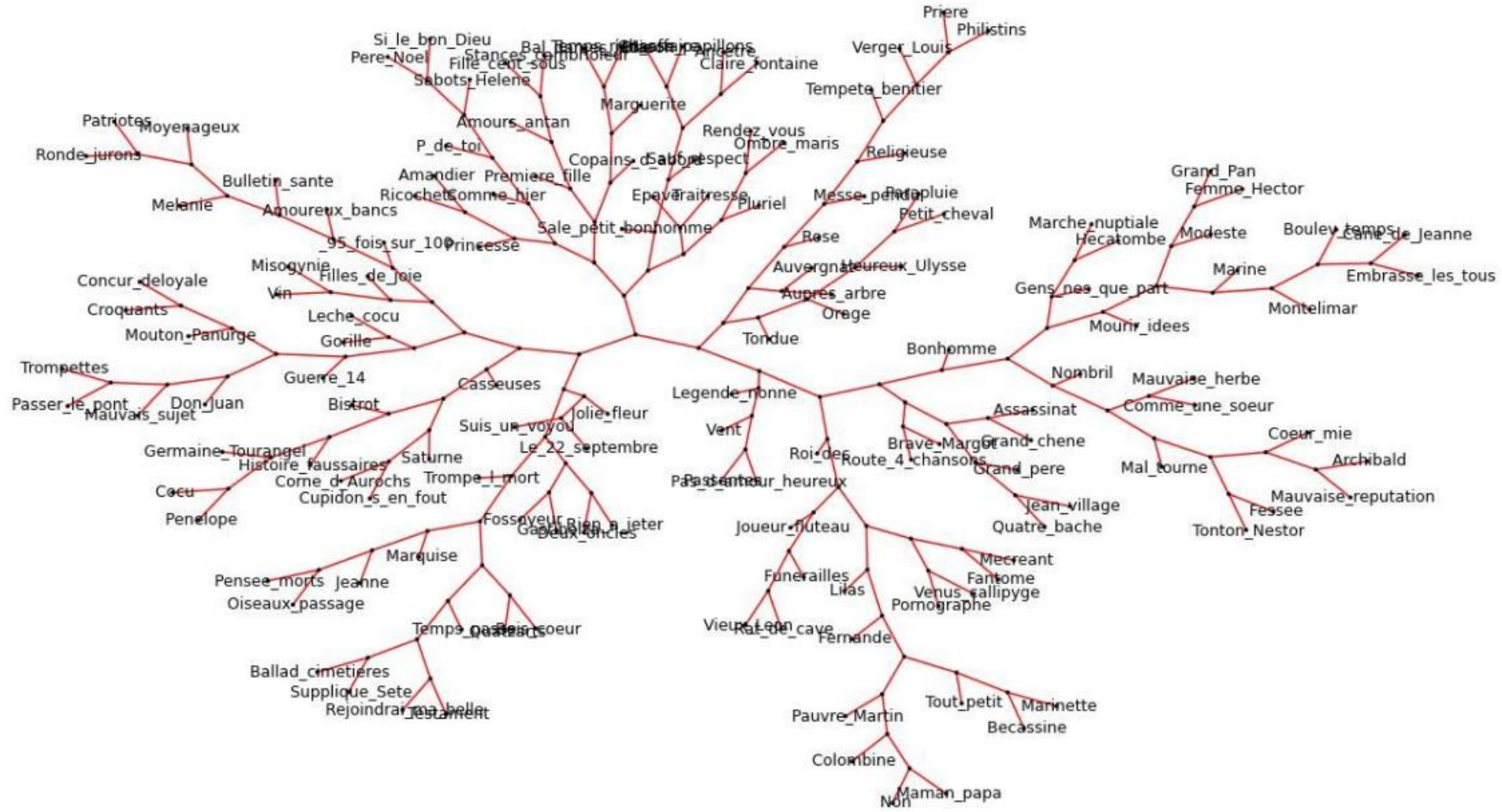
Analyse des correspondances

14 records rom 1953

To 1985

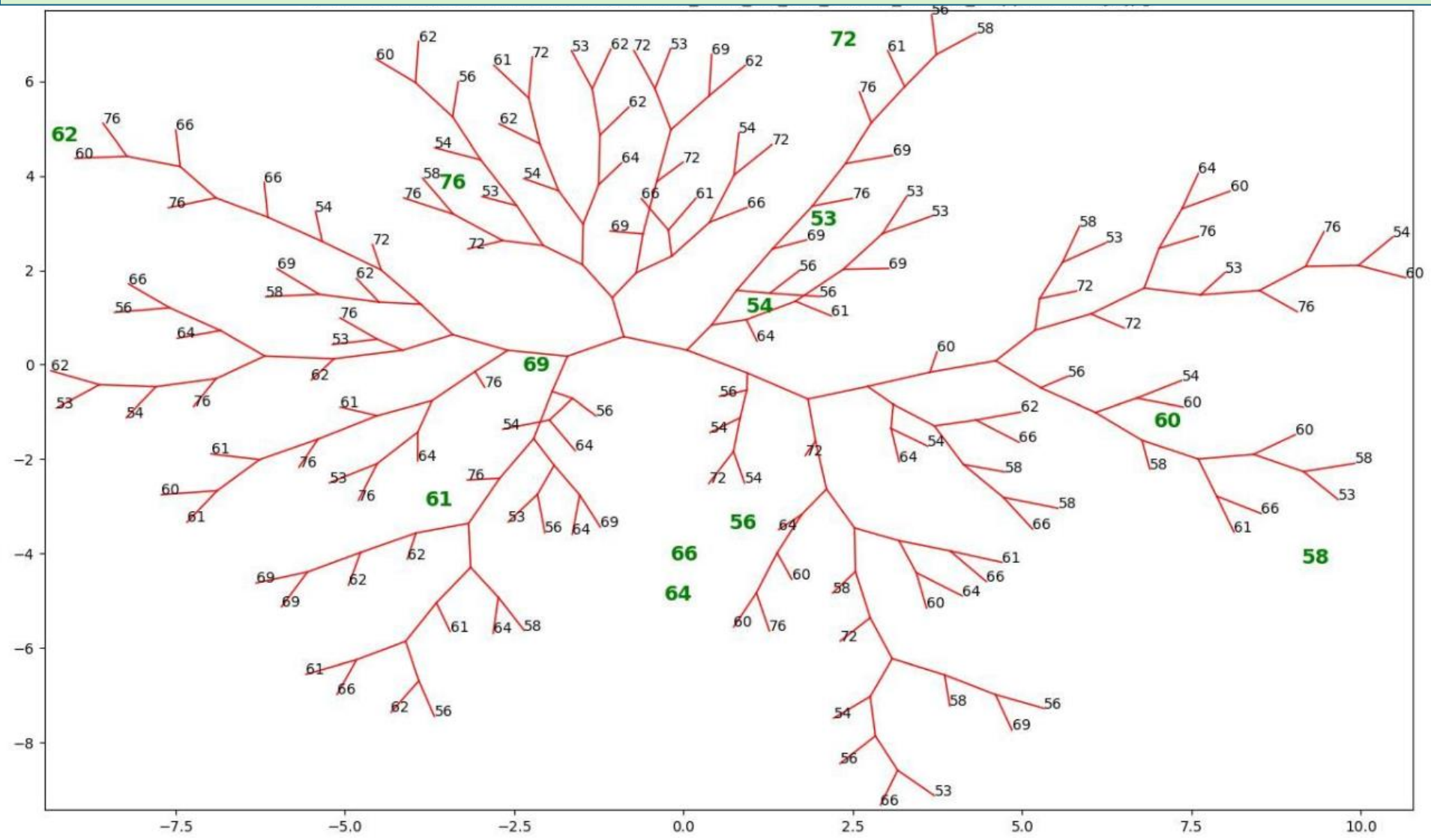
Figure 4. Arbre additif calculé directement sur les 194 chansons

(Figure complémentaire et non incluse dans l'article : « Des outils pour décrire certains corpus de poèmes et de chansons : les arbres additifs simultanés » (Lebart, 2024))



2.1 Corpus Georges Brassens (194 chansons)

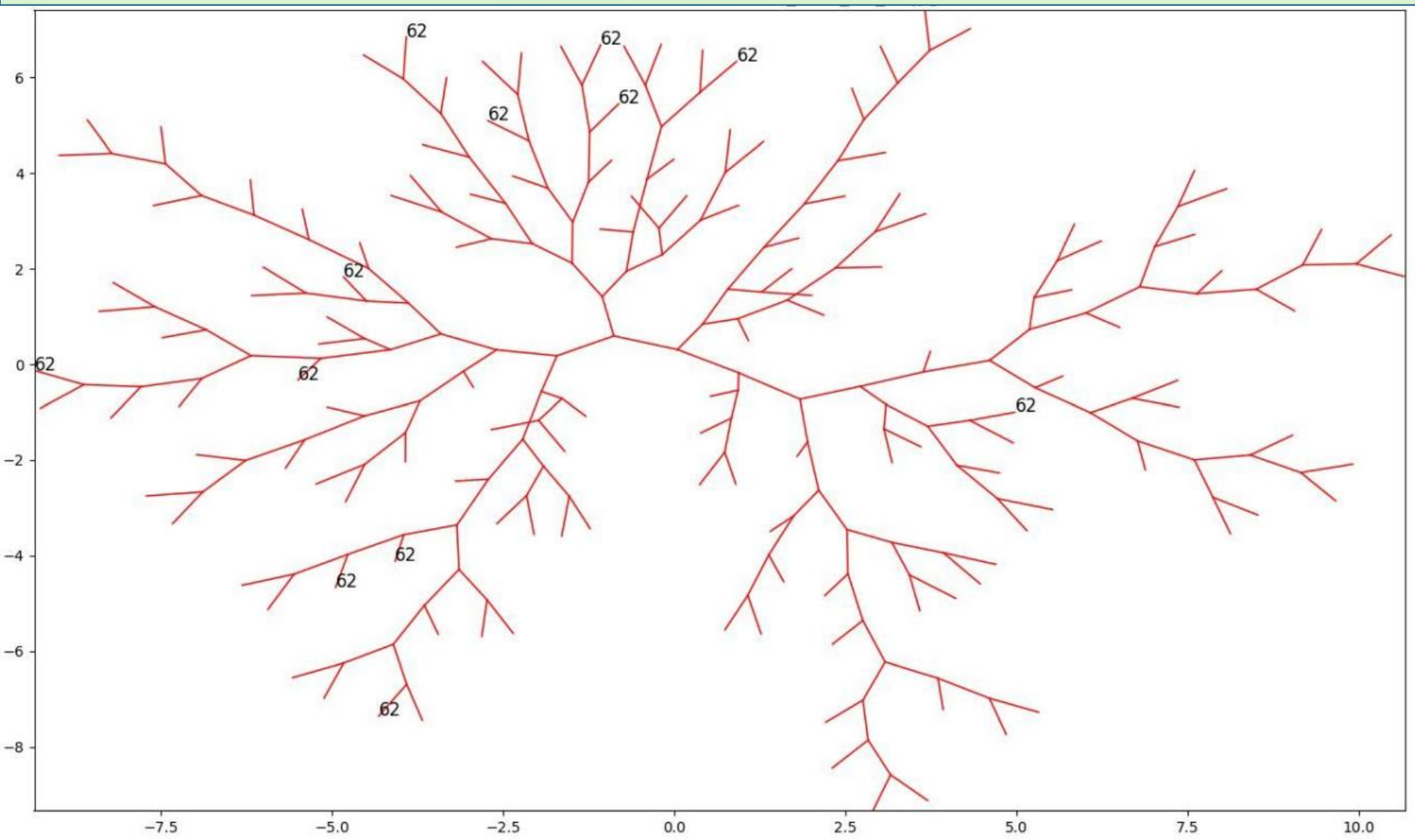
Figure 5. Chaque chanson est repérée par l'année du recueil correspondant, qui figurent également en tant qu'éléments supplémentaires. *(Figure complémentaire et non incluse dans l'article : « Des outils pour décrire certains corpus de poèmes et de chansons : les arbres additifs simultanés » (Lebart, 2024))*



2.1 Corpus Georges Brassens (194 chansons)

Figure 6. Chansons repérées par l'année du recueil, exemple du recueil de **1962**

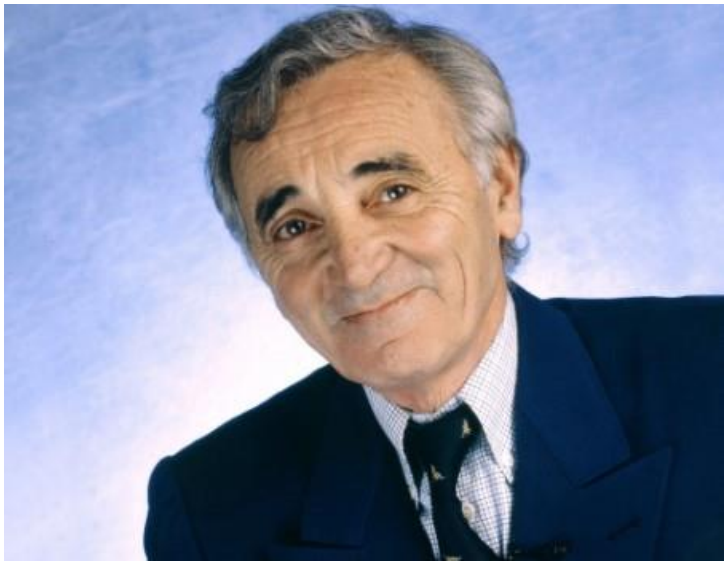
(Figure complémentaire et non incluse dans l'article : « Des outils pour décrire certains corpus de poèmes et de chansons : les arbres additifs simultanés » (Lebart, 2024))



Le corpus Aznavour

Le corpus comprend 295 chansons reparties en 18 recueils (albums). Charles Aznavour (1924 – 2018) a commencé sa carrière musicale dans les années 1940, et a enregistré près de mille deux cents chansons interprétées en plusieurs langues. Il a écrit ou coécrit plus de mille chansons, que ce soit pour lui-même ou d'autres artistes. Elles sont publiées dans le livre : (« Chansons : L'intégral ». Aznavour, 2010).

2.2 Corpus Charles Aznavour (295 chansons, 18 recueils)



Aspect peut-être moins connu du grand public, Aznavour fut longtemps parolier pour d'autres chanteurs, avant d'être un des chanteurs français les plus reconnus en dehors du monde francophone.

Il a écrit pour Édith Piaf, Eddie Constantine, Gilbert Bécaud, Juliette Gréco, Sylvie Vartan, Mireille Mathieu. Il a collaboré avec Fred Astaire, Frank Sinatra, Andrea Bocelli, Bing Crosby, Ray Charles, Liza Minnelli, Tom Jones, etc.

Aznavour était un homme de scène, un grand acteur de cinéma récompensé en tant que tel, un voyageur infatigable, avec des engagements politiques parfois spectaculaires, et des décorations internationales impressionnantes… un peu aux antipodes de Brassens.

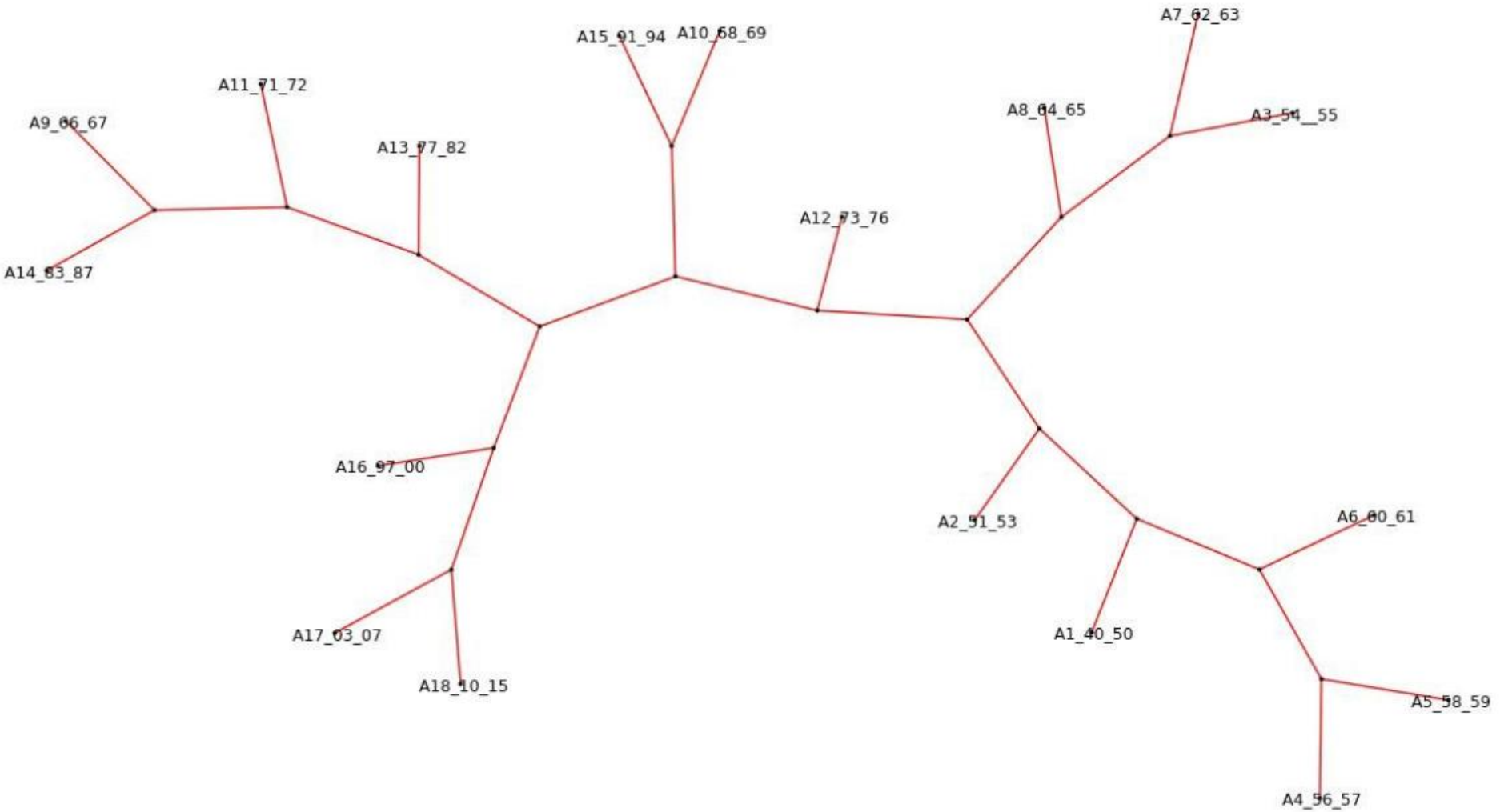
2.2 Corpus Charles Aznavour (295 chansons, 18 recueils)

Figure 7 : Aznavour. Esquisse sommaire du tracé de l'arbre additif avec représentation simultanée des graphies et des textes. Ce sont ici les années récentes qui constituent la partie droite de l'arbre (« 10-15 » signifie « 2010-2015 »). Même limitation concernant le nombre de graphies (6 éléments caractéristiques par texte). Les textes (albums) sont repérés par l'année de parution, de 1940 à 2015 (intervalle de temps deux fois plus large que pour Brassens, et pour Brel qui va suivre). Contrairement à Brassens, le vocabulaire est simple, peu littéraire, mais parfois plus cosmopolite.

2.2 Corpus Charles Aznavour (295 chansons, 18 recueils)

Figure 8 : Esquisse sommaire du tracé de l'arbre additif des recueils. Ce sont ici les années récentes qui constituent la partie droite de l'arbre

(« 10-15 » signifie « 2010-2015 »).



2.2 Corpus Charles Aznavour (295 chansons, 18 recueils)

Figure 9: Représentation simultanée des graphies et des textes (6 graphies caractéristiques par texte). Les textes (albums) sont repérés par l'année de parution, de 1940 à 2015 (intervalle de temps deux fois plus large que pour Brassens, et pour Brel qui va suivre).

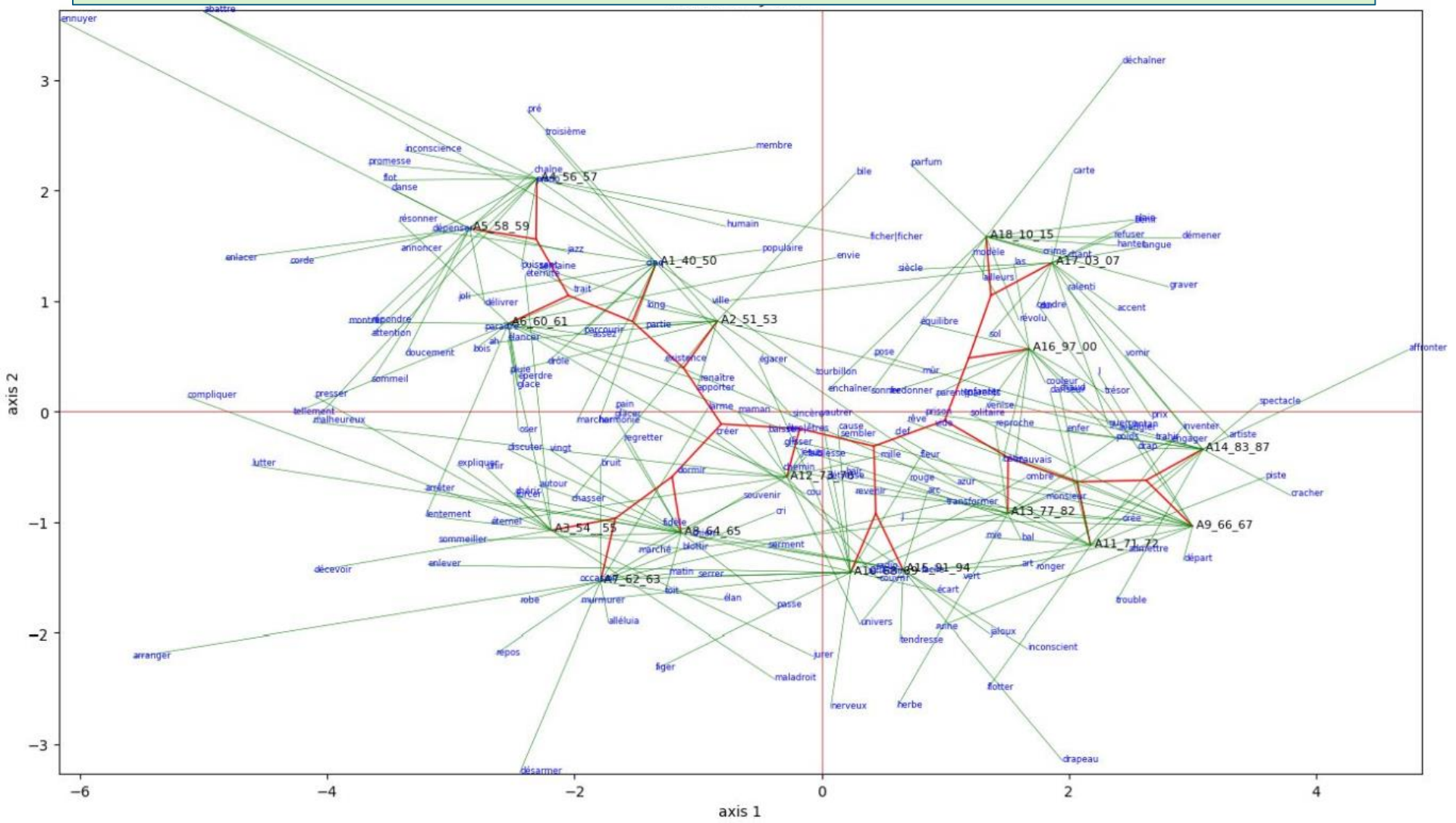


Figure 10. Arbre additif calculé directement sur les 295 chansons

(Figure complémentaire et non incluse dans l'article : « Des outils pour décrire certains corpus de poèmes et de chansons : les arbres additifs simultanés » (Lebart, 2024).)

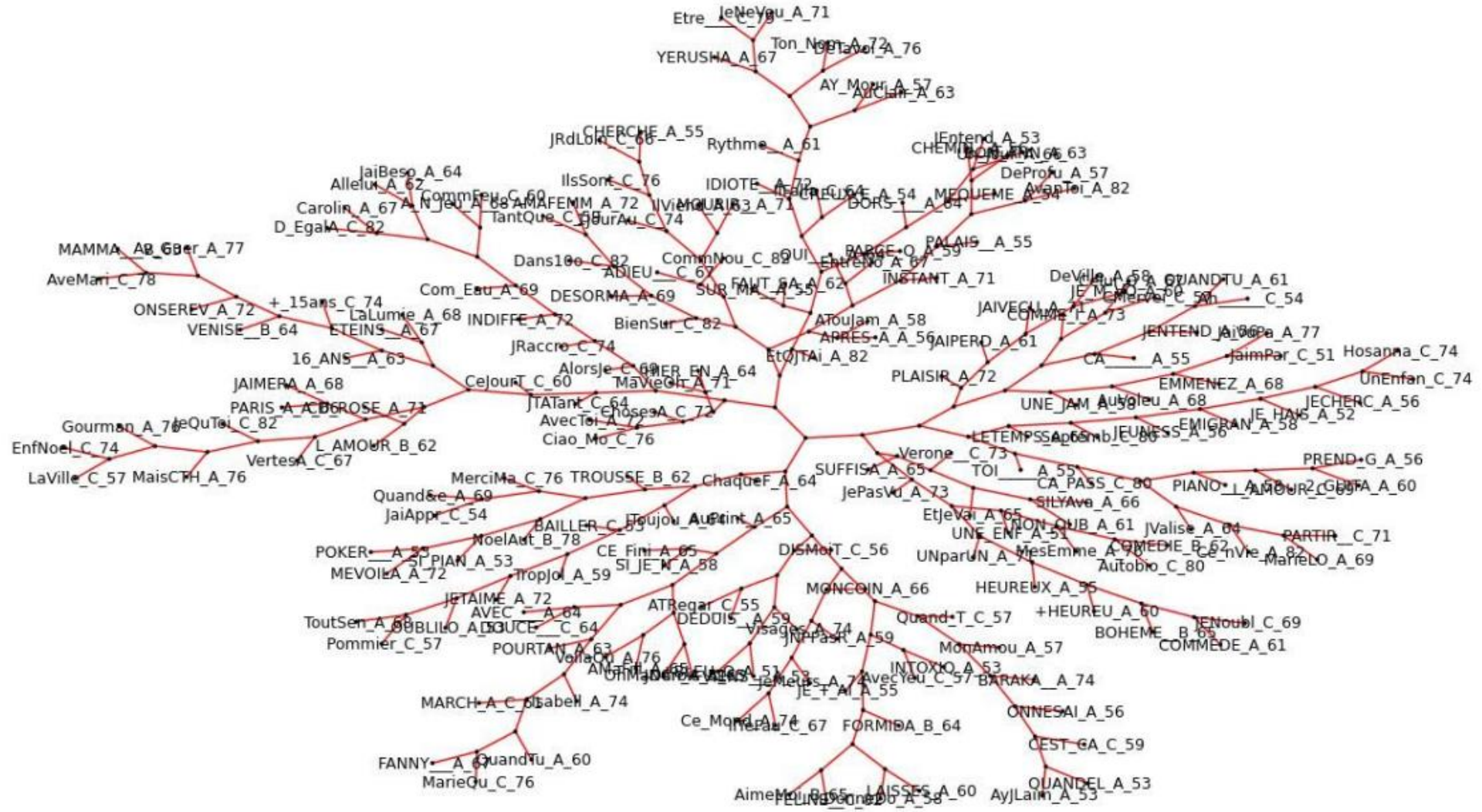
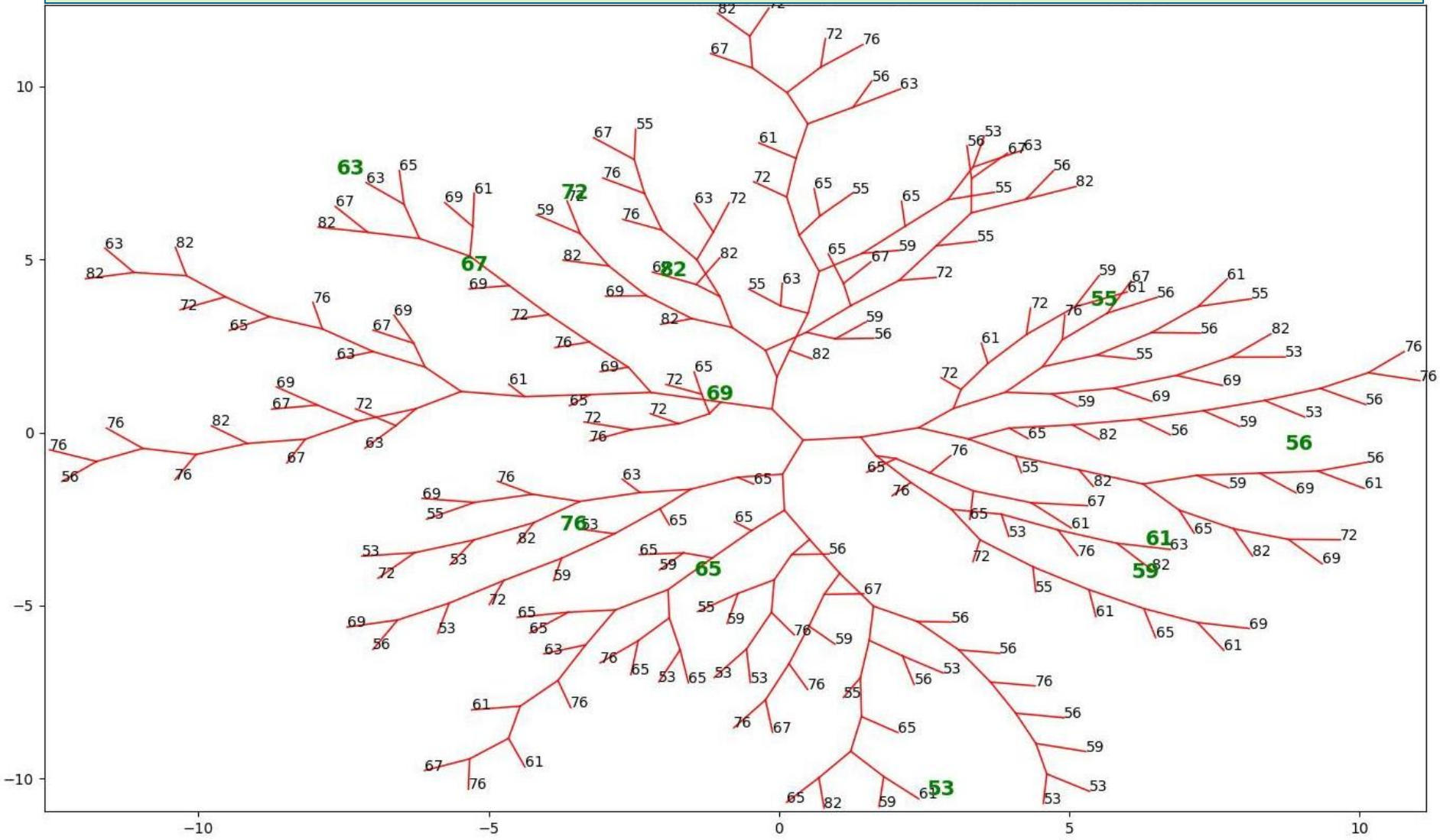


Figure 11. Arbre additif calculé directement sur les 295 chansons décrites ici par l'année du recueil correspondant, ceux-ci figurent également en tant qu'éléments supplémentaires.

(Figure complémentaire et non incluse dans l'article : « Des outils pour décrire certains corpus de poèmes et de chansons : les arbres additifs simultanés » (Lebart, 2024).)



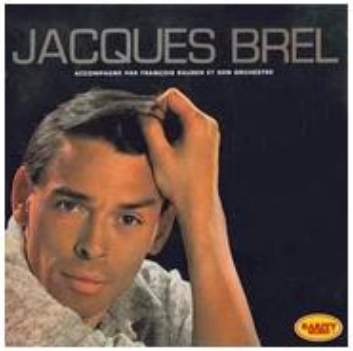
Le corpus Jacques Brel

Jacques Brel (1929 – 1978), auteur-compositeur-interprète, poète, acteur et réalisateur belge est considéré comme un des plus grands auteurs-interprètes de la chanson francophone (plus de 25 millions d'albums vendus). Au sommet de sa popularité, il abandonne pourtant les tours de chant en 1967.

Il fut une source d'inspiration pour bon nombre d'auteurs-interprètes anglophones comme David Bowie, Mort Shuman, Leonard Cohen. Plusieurs de ses chansons sont traduites et chantées par Ray Charles, Nina Simone, Frank Sinatra. Jacques Brel fut une personnalité attachante d'une sensibilité explosive.

1972
Ne Me Quitte Pas

1964
Mathilde



1963
Les Bigotes

1962
Les Bourgeois

1961
N° 5 - Marieke

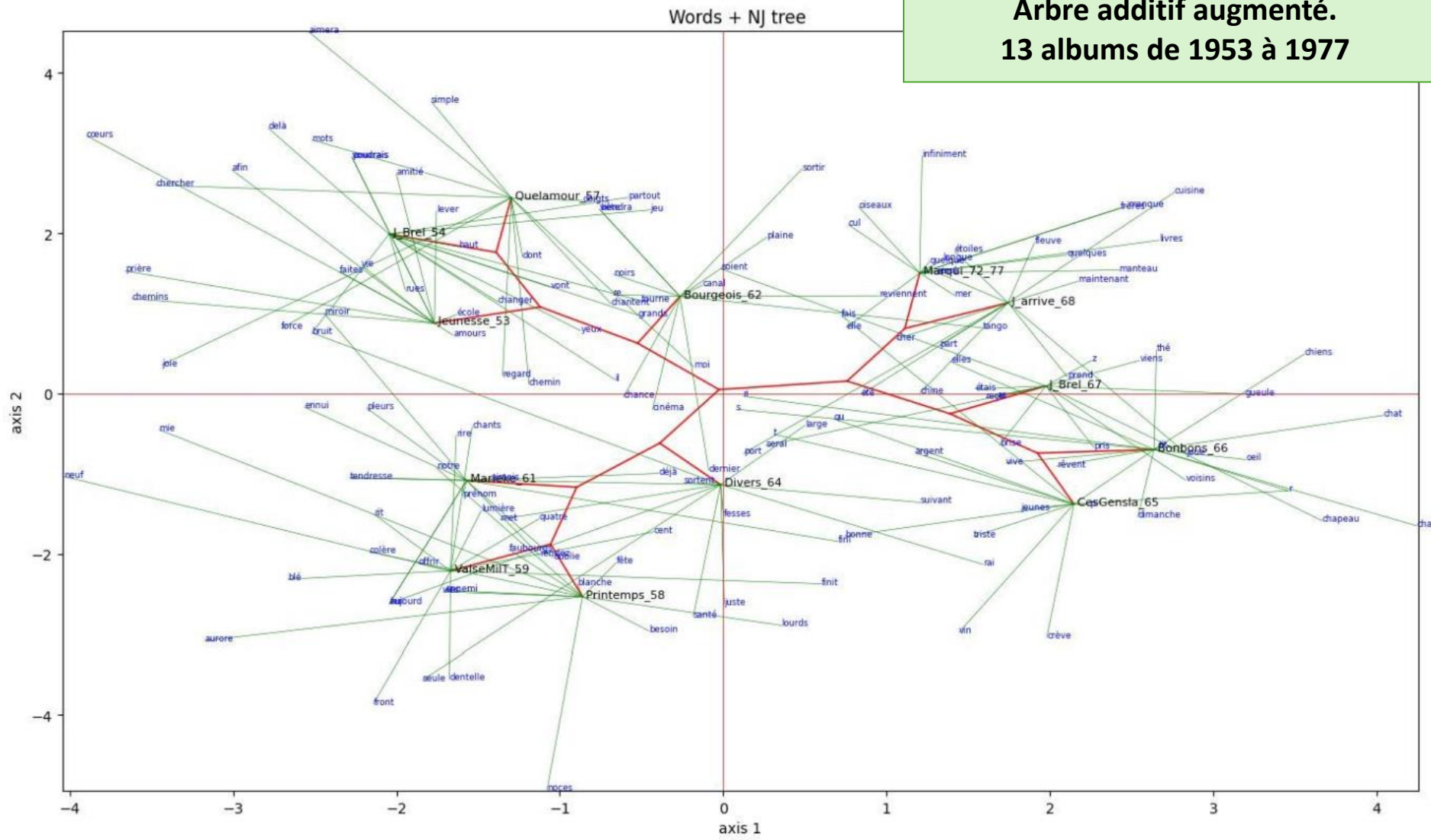
1959
N° 4 - La Valse à Mille Temps

2.3 Corpus Jacques Brel (157 chansons, 13 recueils)

Figure 12

Arbre additif augmenté.

13 albums de 1953 à 1977



2.3 Corpus Jacques Brel (157 chansons, 13 recueils)

Lecture de la figure 12 :

Différentes phases de la vie de Jacques Brel se retrouvent sur ce graphique, avec l'enthousiasme, la gaieté, l'émotion des premières chansons (quadrant haut gauche). Une certaine aigreur et des critiques parfois d'une grande férocité (quadrant bas droit).

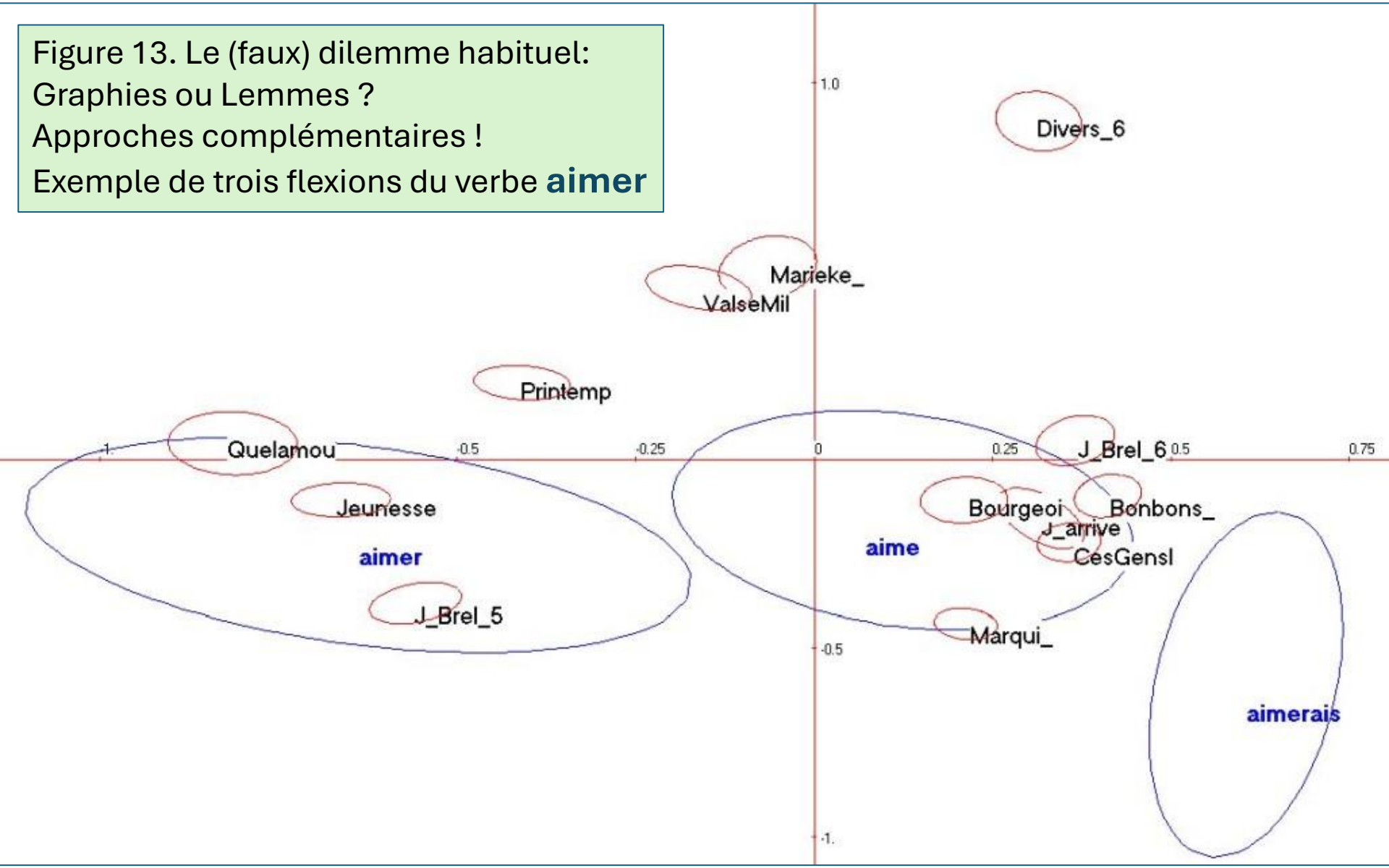
La position détachée à droite des formes graphiques *chat, chats, chiens*, mais aussi *thé, vin* caractérisent l'époque des railleries vis-à-vis de la tiédeur des modes de vie sédentaires et bourgeois, loin des *aimera, amours, simple, chemins* des chansons de jeunesse, à gauche.

Mais ces 13 recueils de 157 chansons sont eux-mêmes hétérogènes, et les thèmes cités peuvent aussi cohabiter à l'intérieur même d'un recueil. L'esquisse d'interprétation précédente concerne seulement les thèmes dominants dans les recueils.

Une analyse directe des 157 chansons montre encore ici la variabilité à l'intérieur des recueils.

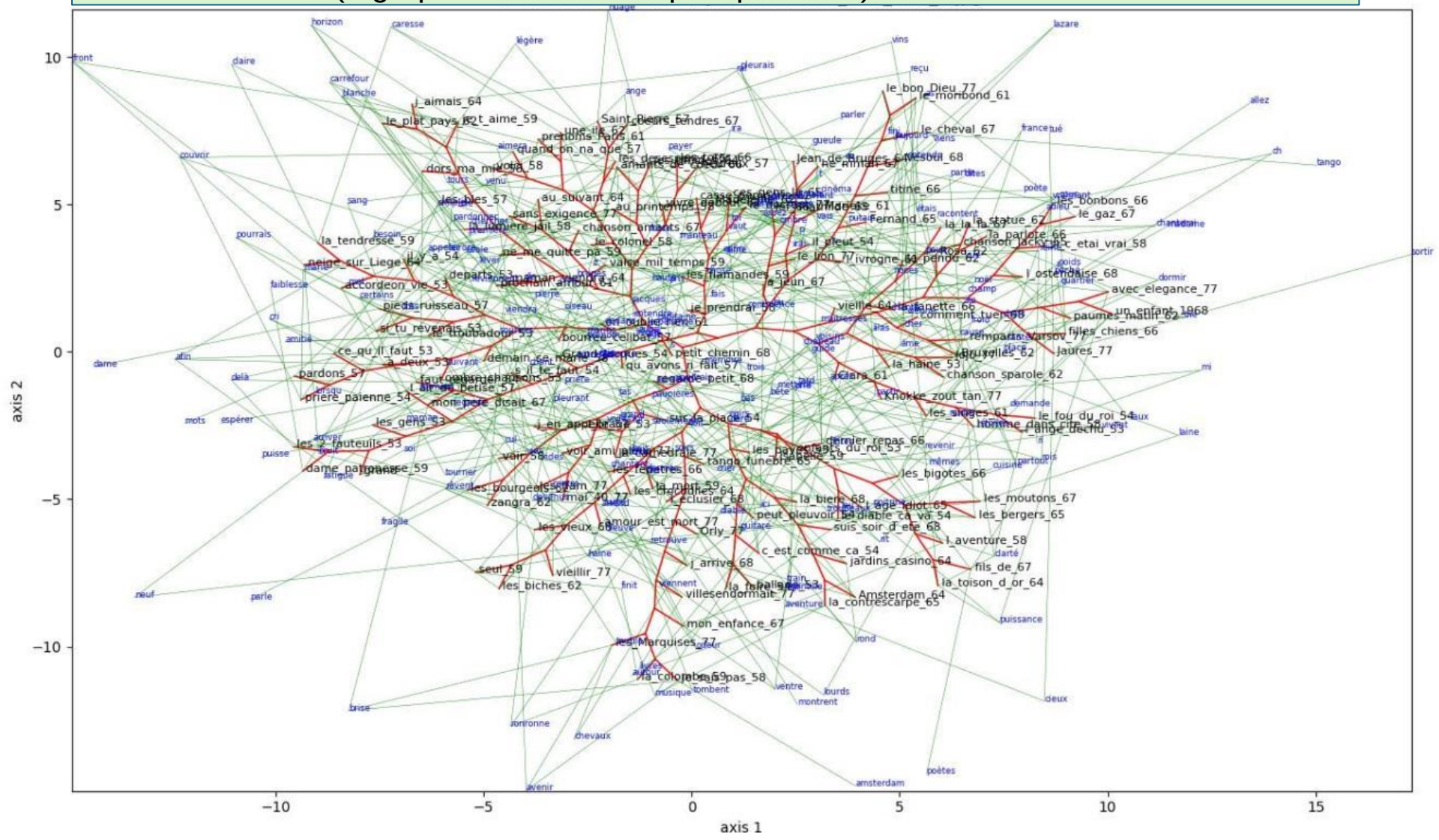
2.3 Corpus Jacques Brel (157 chansons, 13 recueils)

Figure 13. Le (faux) dilemme habituel:
Graphies ou Lemmes ?
Approches complémentaires !
Exemple de trois flexions du verbe **aimer**



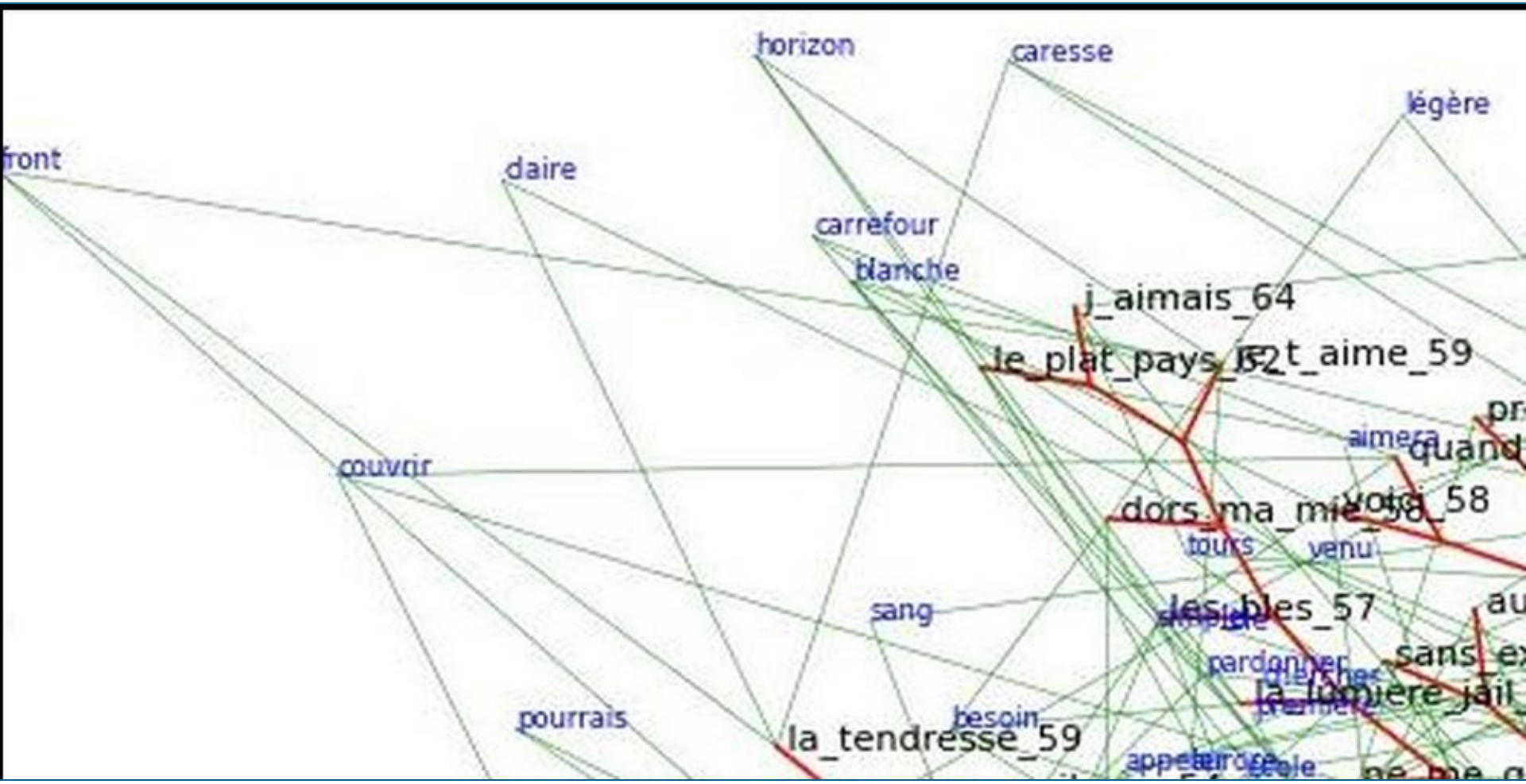
2.3 Corpus Jacques Brel (157 chansons, 13 recueils)

Figure 14: Apparemment inextricable représentation simultanée des graphies et des 157 chansons (3 graphies caractéristiques par texte).



2.3 Corpus Jacques Brel (157 chansons, 13 recueils)

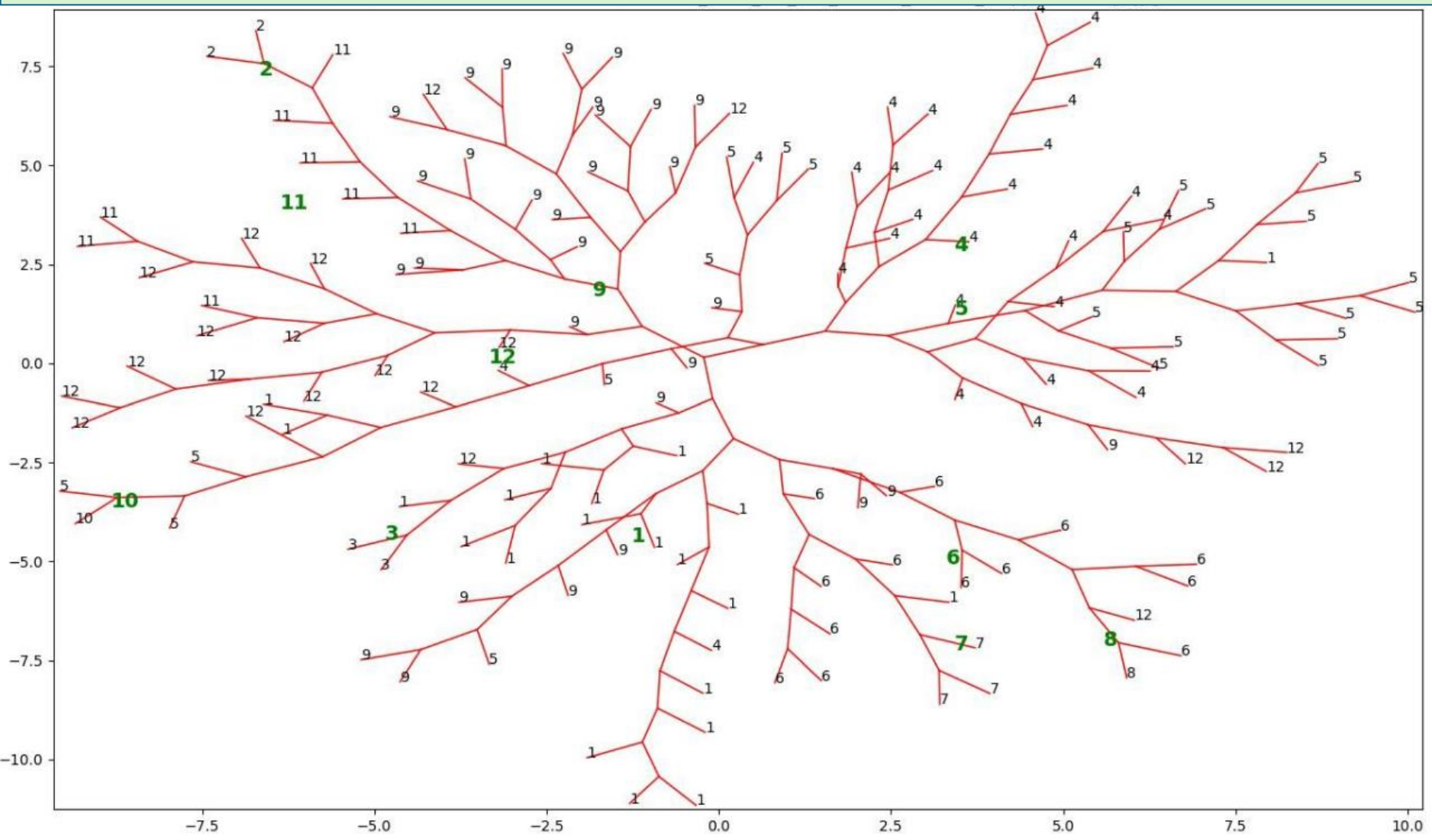
Figure 15: Exemple de zoom pour la représentation simultanée des graphies et des 157 chansons .



2.3 Corpus Jacques Brel (157 chansons, 13 recueils)

Figure 16. Chaque chanson est repérée par le numéro de l'album. Ces numéros figurent également en tant qu'éléments supplémentaires

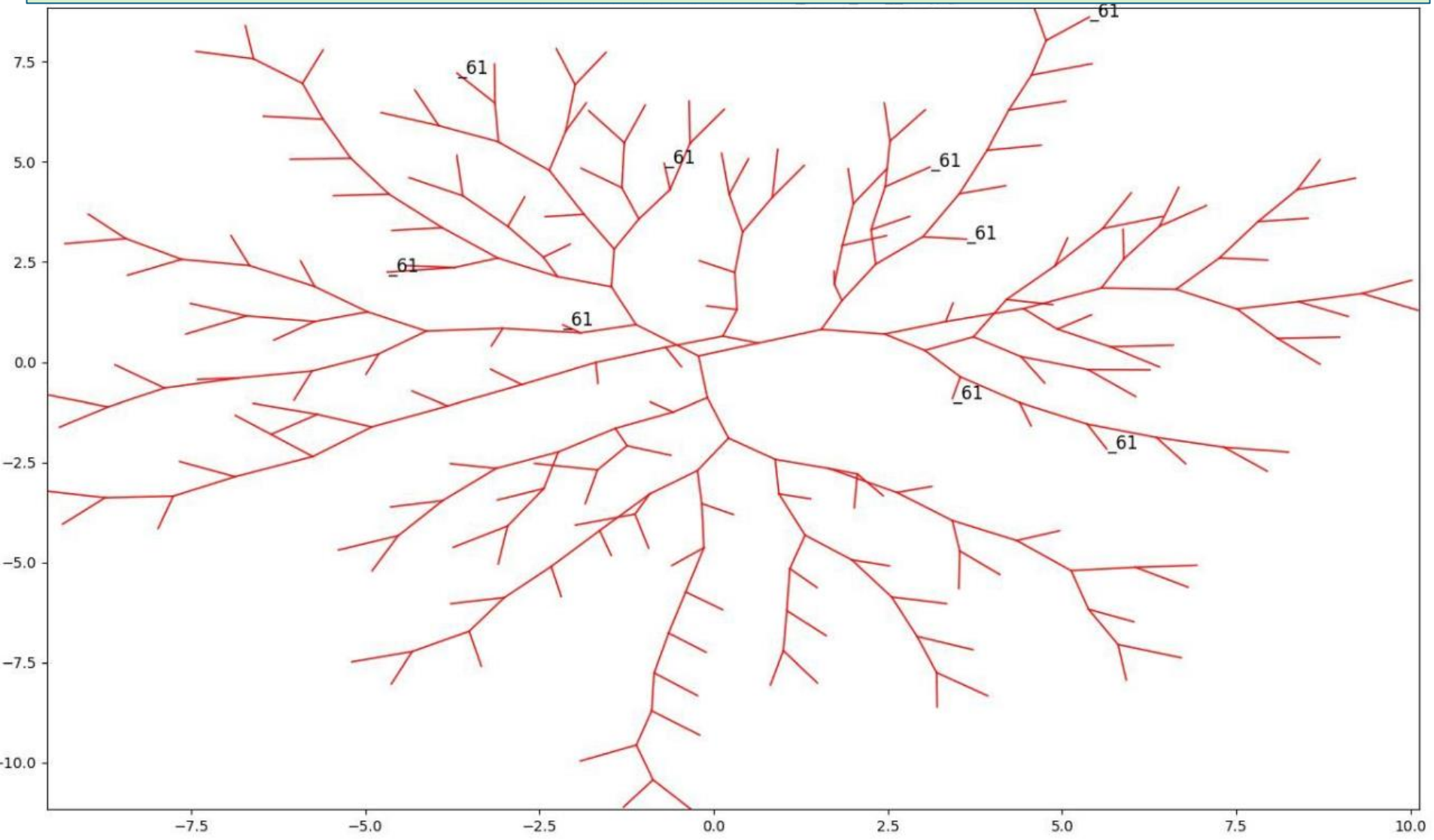
(Figure complémentaire et non incluse dans l'article : « Des outils pour décrire certains corpus de poèmes et de chansons : les arbres additifs simultanés » (Lebart, 2024).)



2.3 Corpus Jacques Brel (157 chansons, 13 recueils)

Figure 17. Chansons repérées par l'année du recueil, exemple du recueil de **1961.**

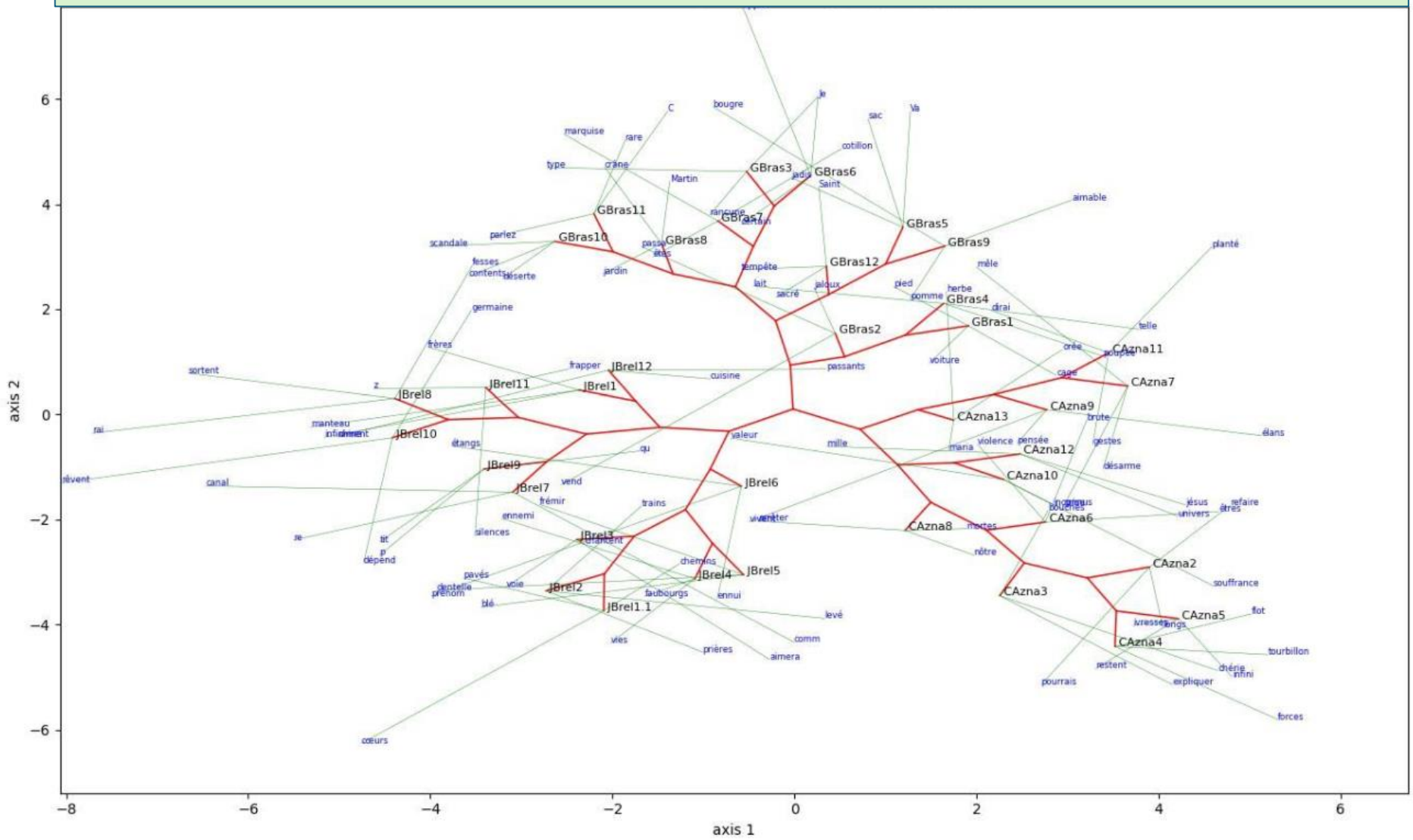
(Figure complémentaire et non incluse dans l'article : « Des outils pour décrire certains corpus de poèmes et de chansons : les arbres additifs simultanés » (Lebart, 2024))



2.4 Corpus Global (1950- 1982)

Complément non inclus dans la communication (Lebart – 2024)

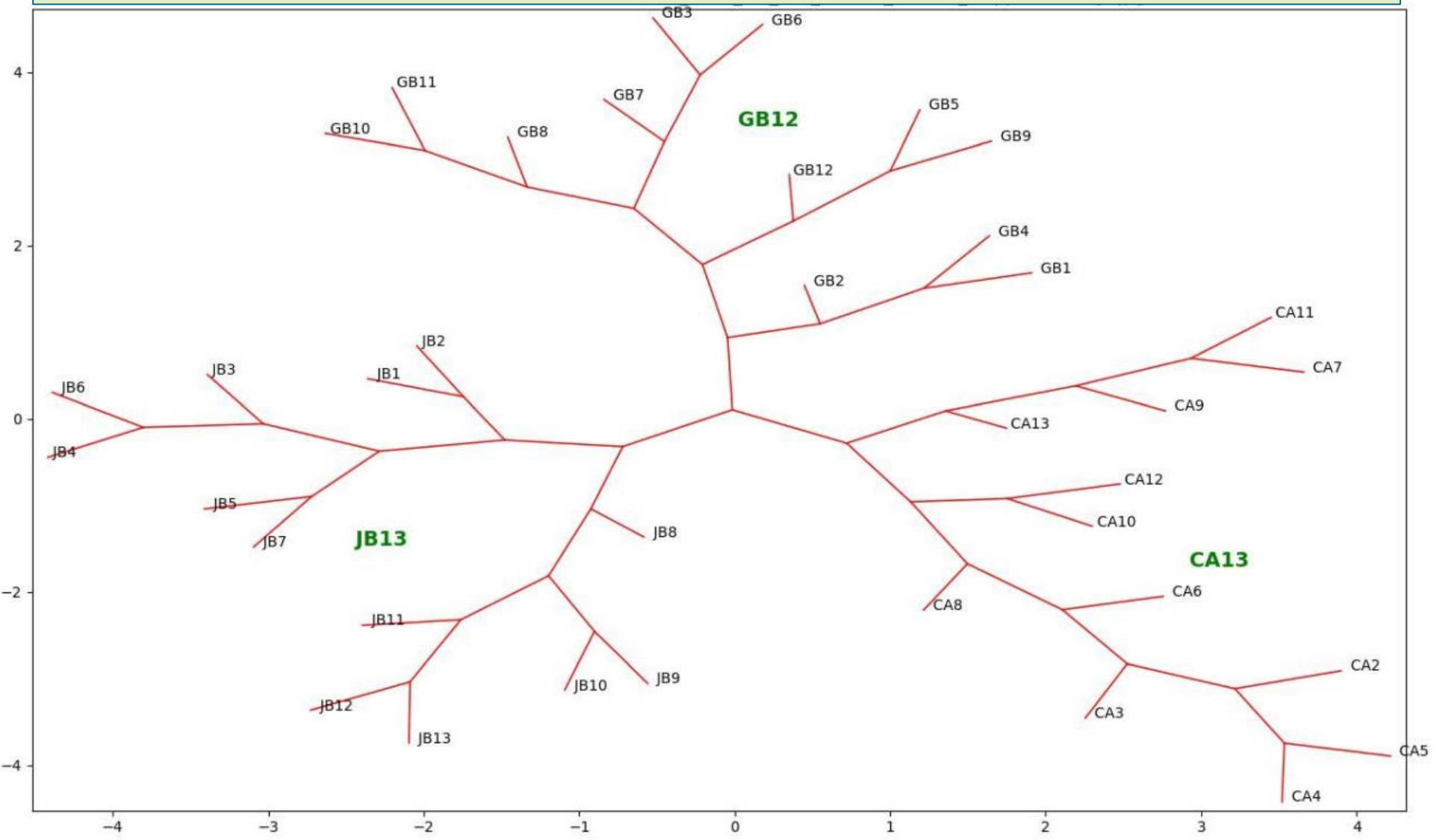
Figure 18: Analyse globale des recueils des trois auteurs par arbre additif simultané. Séparation parfaite des trois auteurs.



2.4 Corpus Global (1950- 1982)

Complément non inclus dans la communication (Lebart – 2024)

Figure 19: Analyse globale des recueils des trois auteurs par arbre additif simultané. Séparation parfaite des trois auteurs. Auteurs en tant qu'éléments supplémentaires.



2.4 Corpus global et complement : table lexicale classique: “Inaugural address” corpus

Complément non inclus dans la communication (Lebart – 2024) et présenté à la conférence “Data Science and Social Research”, Naples, 2024.

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].

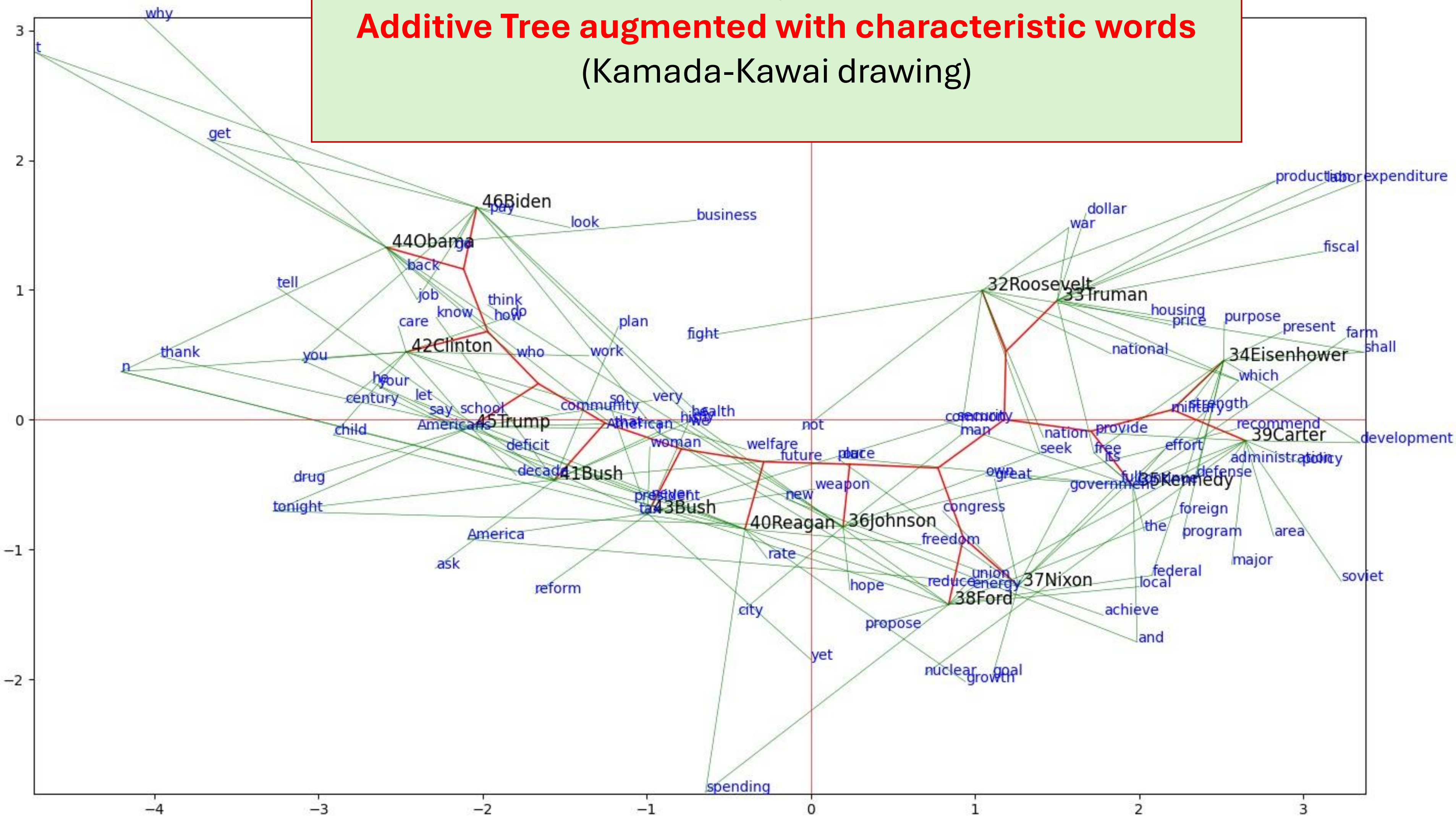
As a check, the corpus was also lemmatized using the software TreeTagger (Schmid, 1994), with elimination of function words and prepositions.

Complément pour une table lexicale classique: “Inaugural address” corpus

State of the Union speeches 1942- 2024

Additive Tree augmented with characteristic words

(Kamada-Kawai drawing)



Conclusion

Comme en analyse des correspondances avec ses représentations simultanées, le **pattern observable des points-colonnes** (textes, recueils) nous dit ***comment*** ces points s'organisent, et la **présence des point-lignes (graphies)** nous disent ***pourquoi*** ils se sont organisés de cette façon : des textes sont proches parce qu'ils utilisent souvent les mêmes mots.

Et les mots habillent de chair lexicale le squelette de l'arbre additif.

Conclusion

Les méthodes de visualisation de données utilisent certes des méthodes algébriques ou algorithmiques voisines de celles de l'intelligence artificielle. On a pu en effet montrer que l'AC est aussi une méthode neuronale (Lebart, 1997).

Mais une visualisation n'est pas une décision à prendre, ni une tâche à exécuter. C'est presque le contraire. On ne pose pas de question pour agir, on soumet des données pour comprendre et réfléchir.

L'approche est non-supervisée, une phase de travail dont le *Deep Learning* aura d'ailleurs de plus en plus besoin selon les prédictions du texte de référence de Le Cun *et al*. (2015).

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…)

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

$\mathbf{D}_p$ 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".**

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).

Data visualization methods certainly use algebraic or algorithmic methods similar to those of artificial intelligence. In some respect, CA is also a neural 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.

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.

Bibliographie

Aznavour C. (2010). *Chansons : l'intégrale*. Points. Don Quichotte Editions. Paris : Seuil.

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.

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., 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.

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

Gracias

Grazie

Obrigado

Merci

Danke

Choukrane

55