MAP Syntax diagrams : visualiser, décrire et comparer les séquences onomastiques divines de l'Antiquité

Sébastien Plutniak

MAP Syntax diagrams : visualiser, décrire et comparer les séquences onomastiques divines de l'Antiquité

Sébastien Plutniak

Table des matières

1 Introduction 2

2 Utilisation de l'application 3
- 2.1 Chargement des données . . . 3
- 2.2 Sélection des données . . . 4
- 2.3 Production des résultats descriptifs . . . 4
  - 2.3.1 Longueur des séquences . . . 4
  - 2.3.2 Statistiques descriptives . . . 6
  - 2.3.3 Visualisation des diagrammes syntaxiques . . . 7
  - 2.3.4 Définir des sous-ensembles de séquences . . . 7
  - 2.3.5 Étudier la position syntaxique de certains éléments . 7
- 2.4 Fonctions de comparaison . . . 7
  - 2.4.1 Visualisation des différences . . . 7
  - 2.4.2 Mesure des différences . . . 8
- 2.5 Export des résultats . . . 8

3 Méthodes 9
- 3.1 La fréquence des transitions entre symboles . . . 9
- 3.2 Les graphes . . . 10
- 3.3 Mesure de distance entre graphes . . . 10

4 Lecture des diagrammes 11
- 4.1 Diagramme des séquences . . . 11
- 4.2 Diagramme de comparaison . . . 14

5 Interprétation des résultats numériques 15
- 5.1 Statistiques descriptives individuelles des diagrammes . . . 15
  - 5.1.1 Propriétés basiques . . . 15
  - 5.1.2 Propriétés du premier élément mis en évidence . . . 15
  - 5.1.3 Richesse globale des types de séquence . . . 16
  - 5.1.4 Indices de la spécificité des types de séquences . . . 16
  - 5.1.5 Probabilités . . . 17
- 5.2 Distance entre diagrammes . . . 17

6 Conclusion 18

1 Introduction

L'application *MAP Syntax diagrams* [1] permet de représenter graphiquement, de décrire numériquement et de comparer des ensembles de séquences onomastiques divines encodées au format développée par le projet de recherche *Mapping Ancient Polytheims* [2]. Dans ce format, une séquence onomastique en sémitique telle que

l'štrt l'dny l'šmn

sera représentée comme :

Astarté / [mon Seigneur # Eshmoun]

et encodée dans la base de données par une chaîne de caractères telle que :

1 / [2 # 3]

Une séquence en grec telle que :

Ἀφροδίτης καὶ Διὸς Πολιέος καὶ Ἥρας

sera représentée comme :

Aphrodite + [Zeus # Polieus] + Hera

et encodée dans la base de données par une chaîne de caractères suivante telle que :

1 + [2 # 3] + 4

Les nombres font référence aux éléments de la séquence [3], quatre symboles (+, /, #, =) qualifient leurs relations, et des groupements d'éléments sont possibles en utilisant des paires de crochets et de parenthèses [4].

MAP Syntax diagrams est une interface graphique, réalisée grâce aux fonctionnalités du paquet *Shiny* [5] pour R (Chang *et al.* 2019), permettant l'exécution commode d'une série de fonctions [6] principalement destinées à étudier la **syntaxe** des séquences onomastiques MAP. L'interface facilite l'**exploration** des données, en permettant de varier rapidement différents paramètres et d'obtenir rapidement des visualisations et des descriptions numériques de sous-ensembles de séquences, ainsi que de mesures de distances entre les différents sous-ensembles considérés. Elle permet ainsi d'étudier la manière dont sont structurés des ensembles de séquences relatifs à des périodes chronologiques, des espaces où à tout autre critère choisi par l'utilisateur.

1. Accessible à l'adresse `https://splutniak.shinyapps.io/syntax-diagram`.
2. `https://map-polytheisms.huma-num.fr`.
3. Ce sont en fait les identifiants des éléments dans la base de données qui sont employés.
4. Pour une présentation plus détaillée, voir Lebreton et Bonnet 2019.
5. `https://shiny.rstudio.com`.
6. Le code source de l'application est disponible à l'adresse `https://github.com/sebastien-plutniak/map-syntax-diagrams`.

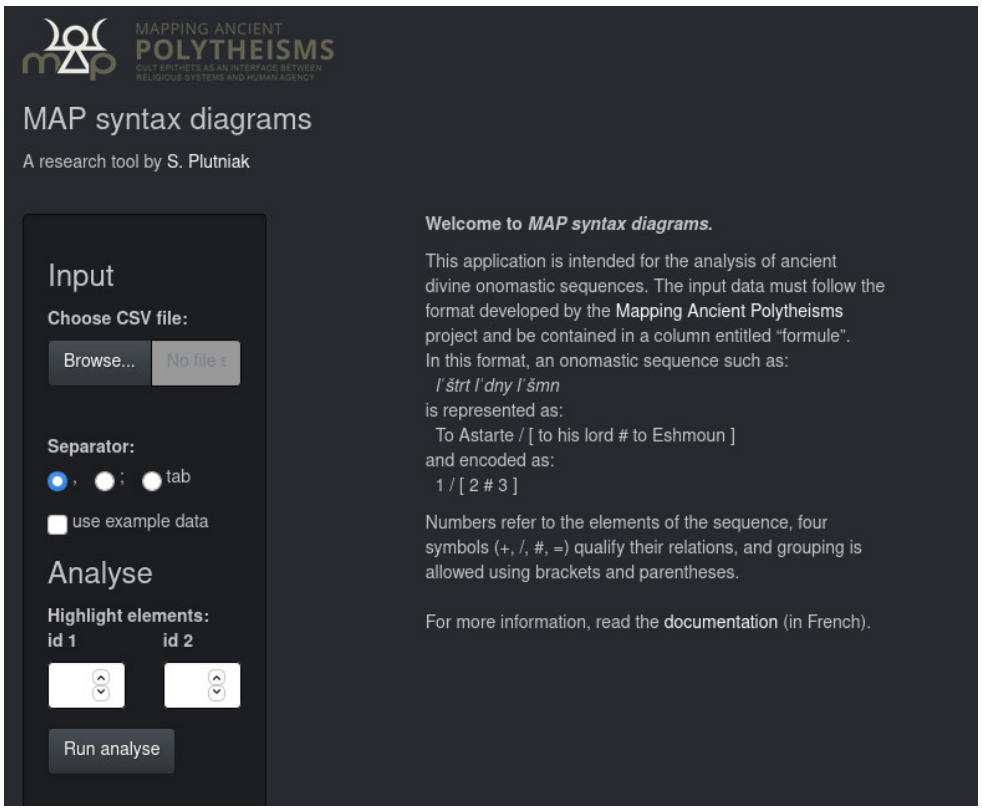

FIGURE 1 – Affichage à l'ouverture de l'application. Sont visibles dans le panneau de gauche : dans la rubrique « *Input* », le sélecteur de fichier csv et le sélecteur de séparateur ; dans la rubrique « *Analyse* », les champs permettant de saisir les identifiants de deux éléments que l'on souhaite étudier plus particulièrement.

2 Utilisation de l'application

L'application est organisée en deux espaces : une **barre latérale** à gauche, où l'utilisateur peut agir pour paramétrer le type d'analyse et de résultats qu'il souhaite obtenir, et un **panneau principal** où sont présentés les résultats (Figure 1). Deux types de résultats seront distingués dans ce qui suit : ceux, dits « descriptifs », qui se limitent à caractériser un diagramme à partir de simples décomptes ou de rapports entre décomptes des propriétés de ses sommets et de ses arêtes ; et ceux dits « comparatifs », dont la production implique des choix de méthode (celui de tel algorithme plutôt qu'un autre).

2.1 Chargement des données

L'application requiert un tableau de données au format csv (*Comma Separated Values*). Si le séparateur est autre que le symbole virgule, il est possible de l'indiquer en sélectionnant le bouton correspondant. Une fois le bon séparateur sélectionné, le tableau s'affiche normalement dans le panneau principal de l'application, sous l'intitulé « *Input table* ». Le tableau doit im-

FIGURE 2 – Message d'erreur affiché si le tableau donné en entrée ne contient pas de colonne correctement intitulée pour les séquences.

pérativement contenir une colonne intitulée « formule » [7], contenant les séquences onomastiques telle qu'elles sont enregistrées dans la base de données MAP. En l'absence d'une colonne intitulée de cette manière, un message s'affichera en bas à droite du panneau principal (Figure 1). Pour explorer les possibilités de l'application, il est également possible d'utiliser l'échantillon de données intégré, en cliquant sur « *use example data* ».

2.2 Sélection des données

Une fois les données chargées, l'utilisateur peut définir l'ensemble, ou le ensembles, de séquences qu'il souhaite étudier et comparer. Il peut aussi bien s'agir de séquences onomastiques relatives à des époque différentes, à des espaces géographiques différents, à des divinités, des contextes épigraphiques différents : en somme, toutes les variables associées aux séquences qu'il est possible d'extraire dans la base de données MAP.

Le menu déroulant « *variable* » permet de sélectionner l'une des colonnes du tableau, qui servira à discriminer les séquences en différents sous-ensembles. Par défaut, la première colonne du tableau est sélectionnée. En dessous du menu déroulant des variables apparaît la liste des modalités de la variable sélectionnée : chaque item est accompagné d'une case qu'il est possible de cocher ou non selon si l'on souhaite ou non retenir cette modalité (Figure 3).

Notons qu'il est possible de ne pas distinguer de sous-ensembles : cela est même recommandé dans un premier temps afin d'examiner les caractéristiques générales des séquences chargées dans l'application. Pour cela, il suffit de cocher le premier item de la liste de modalités, intitulé « *All* ». Suite à cela, un clic sur le bouton « *Run analyse* » lancera l'exécution des calculs.

2.3 Production des résultats descriptifs

2.3.1 Longueur des séquences

Dans le panneau principal, un graphique représentant la **distribution des longueurs** des séquences apparaîtra sous l'intitulé « *Sequence lenghts* ». Ce graphique donne, en abscisse, les longueurs (en nombre de symboles) des séquences, et en ordonnées, le nombre de séquences pour chaque longueur. Notons que les deux axes présentent une échelle logarithmique, afin de faciliter l'affichage de distribution souvent très concentrées dans les valeurs

7. Les intitulés de colonnes « formules », « formula », « FORMULE », et « FORMULES », « FORMULA », sont aussi admis.

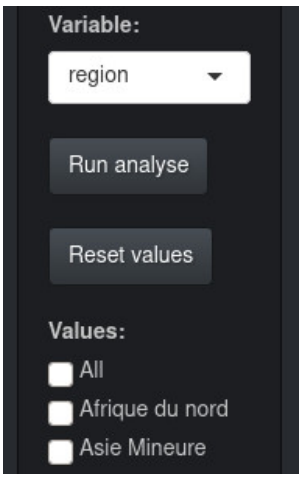

FIGURE 3 – Champs du panneau latéral permettant de sélectionner la variable et ses modalités voulues.

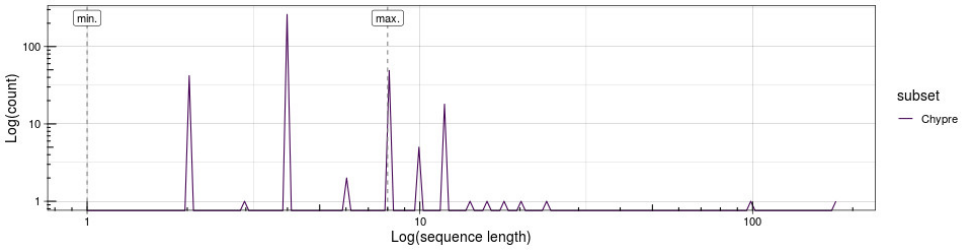

FIGURE 4 – Graphique de la distribution des longueurs des séquences dans les ensembles sélectionnés.

faibles (Figure 4). Ce graphique a été introduit dans l'application non comme pour fournir des résultats, mais plutôt un moyen d'apprécier rapidement la distribution des longueurs et, par conséquent, la manière la plus appropriée de segmenter l'ensemble de données chargées. En effet, imaginons deux cas de figures différents : un premier, où la grande majorité des séquences a une longueur inférieure à 10 symboles, et un deuxième où les longueurs des séquences se répartissent de manière plus homogènes, avec de longues de séquences de plus 20 symboles assez fréquentes. Ayant pris connaissance de ces caractéristiques, on pourra, dans le premier cas, concentrer l'étude sur les séquences inférieures à 10 symboles et considérer que l'essentiel des données de ce sous-ensemble est ainsi étudié ; dans le second cas, il nous faudra tenir compte d'emblée des séquences plus longues.

Cette information permet donc de préciser les longueurs des séquences que nous souhaitons étudier. Une fois la décision prise, les **longueurs minimales et maximales** peuvent être renseignées dans le panneau latéral, à l'aide des deux curseurs du sélecteur « *Seq. length min/max* ». Les limites sélectionnées sont figurées sur le graphique des longueurs des séquences par deux barres verticales en pointillées, respectivement étiquetées « min » et « max » (Figure 5). Insistons sur le fait qu'il s'agit d'un filtrage préalable aux autres opérations d'analyse : si les valeurs renseignées sont 1 (minimum) et 8

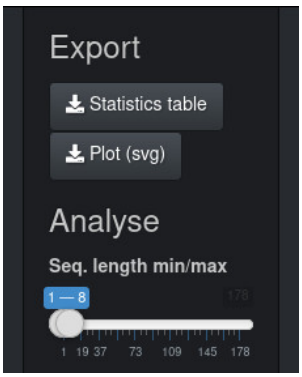

FIGURE 5 – Partie du panneau latéral contenant les boutons d'exportation des résultats ainsi que le sélecteur des longueurs minimales et maximales des séquences à étudier.

Diagram statistics

	Afrique du nord	Asie Mineure	Chypre	Égypte et Nubie	Grèce continentale
Total nr of sequence	397.00	53.00	384.00	259.00	230.00
nr of selected sequences	289.00	50.00	354.00	186.00	216.00
nr of different sequences	10.00	7.00	14.00	20.00	12.00
sequence max. length	7.00	7.00	7.00	7.00	7.00
nr of nodes	24.00	19.00	25.00	29.00	24.00
nr of different symbols	8.00	8.00	8.00	9.00	8.00
nr of terminal symbols	5.00	3.00	3.00	3.00	4.00
nr of articulation points	8.00	5.00	4.00	3.00	4.00
degree centralisation	0.09	0.05	0.12	0.12	0.10
median probability	0.01	0.02	0.03	0.04	0.01

FIGURE 6 – Affichage du tableau des statistiques descriptives dans le panneau central de l'application, ici pour cinq ensembles de séquences.

maximum (ce sont les valeurs par défaut), alors, les séquences de 9 symboles ne seront pas prises en compte dans le reste de l'analyse. De même, si les valeurs 5 et 15 sont renseignées, alors les séquences de longueurs inférieures à 5 ou de longueurs supérieures à 15 ne seront pas prises en compte.

2.3.2 Statistiques descriptives

Sous le graphique des longueurs, après avoir choisi l'ensemble des modalités de la variable sélectionnée (la modalité « *All* »), un tableau est apparu sous l'intitule « *Diagram statistics* » (Figure 6). Ce tableau comporte une **série de mesures** ayant été réalisées sur le sous-ensemble de données définie par l'utilisateur dans le panneau de gauche. Dans le cas présent, il correspond donc aux mesures faites sur l'ensemble des séquences de longueur inférieure ou égale à 8 (paramètre : *min.* = 1, *max.* = 8, *value* = « *All* »). Le détail des mesures est donné ci-dessous, en section 5.1.

2.3.3 Visualisation des diagrammes syntaxiques

Enfin, en deçà du tableau statistique, est également apparu un graphique, sous l'intitulé « *Syntax diagram plot* » (Figure 9). La manière de lire le **diagramme** représenté sera précisé ci-dessous, en section 4.1.

2.3.4 Définir des sous-ensembles de séquences

Après cet examen général des séquences entrées dans l'application, il est possible d'affiner l'analyse en divisant les données en sous-ensembles. Pour cela, la variable souhaitée doit être sélectionnée dans le menu déroulant (Figure 3). La liste des modalités de cette variable se met alors à jour juste en-dessous. Il est dès lors possible de sélectionner les modalités souhaitées. Une fois la sélection effectuée, l'analyse peut être relancée en cliquant sur le bouton « *Run analyse* » Le bouton « *Reset values* » permet de désélectionner en un clic toutes les valeurs ayant été cochées.

2.3.5 Étudier la position syntaxique de certains éléments

L'application permet également d'étudier la position syntaxique de certains éléments particuliers (« éléments » au sens que ce terme prend dans la méthode MAP : il peut autant s'agir de noms que d'épithètes, tous enregistrés dans la base de données). Pour cela, deux champs, situés sous l'intitulé « *Highlight elements* », permettent de saisir les identifiants d'un ou de deux éléments (Figure 1). Si cette information est saisie, alors le tableau statistique et les diagrammes contiendront des informations spécifiques à leur égard (Figure 11).

Remarque importante : si des identifiants d'éléments ont été indiqués, alors la structure du graphe est modifiée par l'introduction de nouveaux sommets. En conséquence, une partie des statistiques descriptives ne sera plus pertinente et ne sera donc pas affichée.

2.4 Fonctions de comparaison

L'objectif de l'application est de faciliter les comparaisons entre séquences relatives à des sous-ensembles. Après avoir sélectionner plusieurs modalités de la variable retenue, l'application recalcule le graphique de distribution de longueur des séquences, le tableau statistique, et le graphique des diagrammes syntaxiques. Chacun des ces éléments contient dès lors les résultats relatifs aux différentes modalités de la variable sélectionnées (par exemple : les séquences onomastiques localisées dans les Îles égéennes et en Grèce continentale, si ces deux modalités de la variable « Grandes régions » ont été sélectionnées).

2.4.1 Visualisation des différences

Les instruments de comparaison sont disponibles dès que deux modalités de la variable retenue sont sélectionnées. D'une part, si exactement deux modalités de la variable sont sélectionnées, une autre case apparaît sous le menu déroulant des variables avec l'intitulé « *Show differences* ». Si cette case est

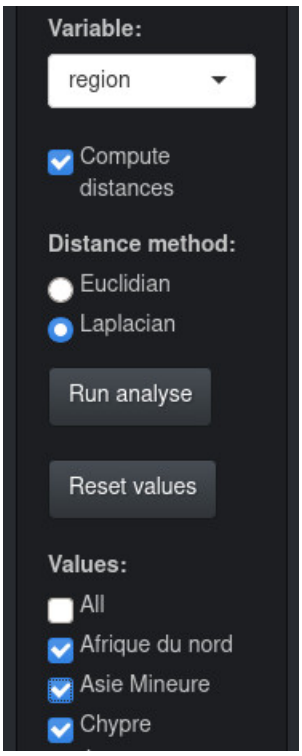

FIGURE 7 – Partie du panneau latéral contenant la case « *Compute distances* » et le sélecteur de méthode.

cochée, l'application affichera un troisième **diagramme**, **mettant en évidence les similitudes** et les différences entre les diagrammes des deux modalités de la variable ayant été sélectionnées. La manière de lire ce diagramme est précisé en section 4.2 (Figure 11). Cette fonction n'est pas disponible lorsque le nombre de modalités sélectionnées est différent de deux.

2.4.2 Mesure des différences

Par ailleurs, une cache à cocher « *Compute distances* » apparaît dans le panneau latéral (Figure 7). Si elle est cochée, un tableau apparaît dans le panneau principal, sous l'intitulé « *Graph distances* », contenant le résultat d'une **mesure de distance** entre les diagrammes des deux sous-ensembles sélectionnés. La manière d'interpréter ces valeurs est précisée ci-dessous, en section 5.2. Lorsque plus de trois modalités de la variable sont sélectionnées, le tableau des distances est complété par la représentation **graphique d'une classification** hiérarchique effectuée sur ces distances (Figure 8.

2.5 Export des résultats

Enfin, on notera la présence dans le panneau latéral gauche, après qu'une analyse ait été exécutée, de deux boutons permettant respectivement le téléchargement du tableau de résultats statistiques (au format csv) et du graphique du diagramme (au format svg) (Figure 5). Notons qu'il est également possible de télécharger le graphique du diagramme au format png à partir

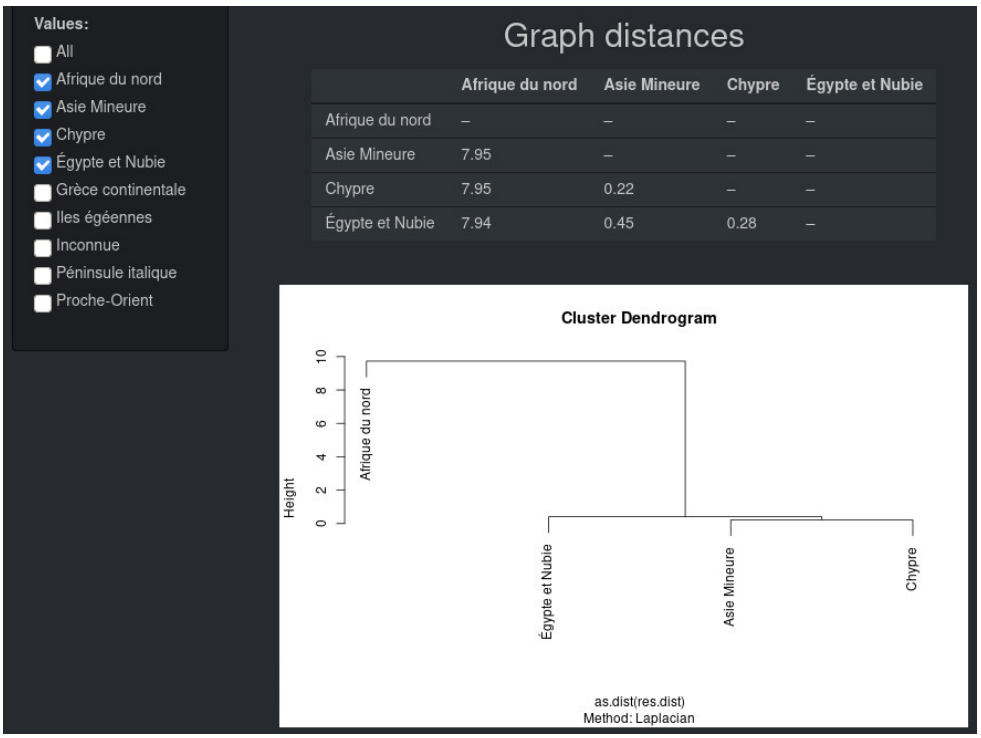

FIGURE 8 – Tableau des distances entre quatre diagrammes, et représentation graphique de la classification hiérarchique correspondante.

du menu déroulant qui apparaît suite à un clic droit sur l'image.

3 Méthodes

3.1 La fréquence des transitions entre symboles

Nous travaillons ici avec des séquences onomastiques dont les « éléments » (au sens MAP, c'est-à-dire les éléments sémantiques en langue naturelle) ont été remplacés par un unique symbole, « x ». L'étude porte donc sur la syntaxe des séquences onomastiques, et non sur leur contenu.

Les séquences onomastiques sont représentées par des séquences à l'aide du paquet *TraMineR* (Gabadinho *et al.* 2011). Une fonction de ce paquet permet de calculer, pour un ensemble de séquences données, les taux de transition entre symboles. Par exemple, considérons la séquence x+x#x. Le taux de transition entre les symboles x et + est de 50%, ainsi que celle entre x et #. La taux de transition entre le symbole + et x est de 100% et celui entre + et # est égal à 0. Ce calcul est généralisé à l'ensemble des paires de symboles pour l'ensemble des séquences et permet d'obtenir une statistique globale des taux de transition.

Il est possible de calculer ces transitions, non pas au niveau de l'ensemble de la séquence, mais à celui de position sur cette séquence. Ceci permet de répondre à une question telle que : sachant que le premier symbole est (par exemple) un « x » quelle est la probabilité que le second symbole soit un « # » ?

Ce calcul a été légèrement adapté aux besoins de *MAP Syntax diagrams* : pour chaque transition entre deux symboles, le nombre de cas attestés est calculé puis rapporté au nombre total de séquences contenues dans l'ensemble considéré. De cette manière, les arêtes du diagramme sont pondérées par la proportion de séquences (par rapport à l'ensemble de séquences considérées) contenant cette transition.

3.2 Les graphes

Les diagrammes employés dans les fonctions et les visualisations de *MAP Syntax diagrams* sont, informatiquement, représentés par des graphes. Leurs manipulation et les calculs qui leurs sont appliqués sont réalisés avec les fonctions du paquet *igraph* (Csárdi et Nepusz 2006).

3.3 Mesure de distance entre graphes

La comparaison de graphes, en particulier de graphes de taille différentes, soulève des difficultés mathématiques importantes, qui ont donné lieu à une très abondante littérature, et non moins abondante variété de méthodes ayant été développée pour déterminer une distance entre deux graphes. En conséquence de cela la comparaison entre plusieurs diagrammes syntaxiques suppose le choix d'une méthode parmi les nombreuses méthodes existantes [8].

Pour la comparaison de deux graphes ou davantage, *MAP Syntax diagrams* procède de la manière suivante :

- les graphes à comparer sont fusionnés dans un nouveau graphe ;
- pour chaque graphe à comparer, un graphe est dérivé du graphe fusionné dont on ne retient que les arêtes présentes dans le graphe à comparer (ceci, de manière à comparer des graphes comportant le même nombre de sommets) ;
- pour chaque paires de graphes dans l'ensemble de graphes à comparer, la matrice d'adjacence pondérée est extraite (la pondération étant la proportion de séquences contenant chaque transition) et une mesure de distance est appliquée ;
- la matrice des distances ainsi obtenue est affichée dans le panneau principal de l'application.

Concernant le choix de la distance, Jurman *et al.* 2010 concluaient en faveur des performances supérieures de la méthode proposée par Ipsen et Mikhailov 2002 ; leur étude ne considérait toutefois que les graphes non pondérés et non orientés. En outre, la méthode d'Ipsen et Mikhailov est particulièrement sophistiquée (fondée sur le spectre du Laplacien des matrices) ; or, une étude récente a montré la validité et la robustesse de méthodes plus simples, appelant à n'employer des procédures plus sophistiquées que sous condition de démontrer leur meilleure performance (Martínez et Chavez 2019). Toutefois, la conclusion de Martínez et Chavez en faveur de l'emploi

8. Pour une panorama sur cette question, et une comparaison de six mesures de distance entre graphes, voir Jurman *et al.* 2010.

de la simple distance euclidienne se fondait elle aussi sur une étude limitée aux graphes non pondérés et non orientés.

Compte tenu du fait que les diagrammes MAP sont orientés et pondérés, le choix a été fait de proposer deux mesures de distance : la distance euclidienne et la *graph diffusion kernel distance* adaptée aux graphes pondérés (Hammond *et al.* 2013). Pour cette dernière, nous utilisons l'implémentation disponible dans le package R *NetworkDistance* (You 2019).

La classification hiérarchique sur la matrice des distances est ensuite réalisée des distances avec l'algorithme Ward 2 (Murtagh 1985).

4 Lecture des diagrammes

L'application peut générer deux types de diagrammes : ceux synthétisant les ensembles de séquences, et ceux résultant de la comparaison de deux diagrammes de premier type.

4.1 Diagramme des séquences

Les diagrammes se lisent de gauche à droite, le premier symbole étant un symbole vide (« Ø ») (Figure 9). Les flèches vers la droite pointent d'abord vers les symboles possibles en première position des séquences onomastiques, puis vers les symboles possibles en deuxième position, *etc.* Pour chaque transition (représentée par les arêtes du diagramme) la proportion arrondie de séquences contenant cette transition, par rapport au nombre total de séquences considérées, est indiquée sur l'arête, en noir. La couleur et l'épaisseur des arêtes font également fonction de cette proportion : plus elle est élevée, plus la couleur tend vers le rouge et plus l'arête est épaisse. Sachant que le nombre total de séquences considérées pour construire le diagramme est indiqué dans le tableau de statistique, il suffit, pour savoir combien de séquences contiennent une transition donnée, de multiplier le nombre total de séquences par la valeur indiquée sur l'arête représentant la transition en question. Par exemple : si le diagramme contient 100 séquences, et que l'arête entre le symbole initial « Ø » et le symbole « [» en position 1 a une pondération de 0.45, on déduira que 100 * 0.45 = 45 séquences débutent par un crochet. Le **diamètre des nœuds** est proportionnel à leur degré pondéré complet (c'est-à-dire, pour un sommet donné, les valeurs de l'ensemble des arêtes qui lui sont reliées). Les nœuds de couleur grisée correspondent à des **symboles terminaux** des séquences (ce sont, en terme de graphes, les « feuilles » de l'arbre). Si des identifiants ont été saisis dans les champs de l'entrée « *Highlight elements* », alors le diagramme sera modifié : les éléments sélectionnés, s'ils existent dans l'ensemble de séquences étudiée, seront représenté par des nœuds de forme quadrangulaire (Figure 10), contenant le label « id1 » ou « id2 » pour, respectivement, les éléments dont l'identifiant a été saisi dans le premier et le deuxième champ de « *Highlight elements* ».

Il est, de cette manière, possible d'étudier de manière synthétique et visuelle la manière dont sont construites un ensemble de séquences onomastiques. Ceci permet d'identifier efficacement les séquences les plus fréquentes (en suivant les arêtes les larges et vives) et de détecter les cas plus

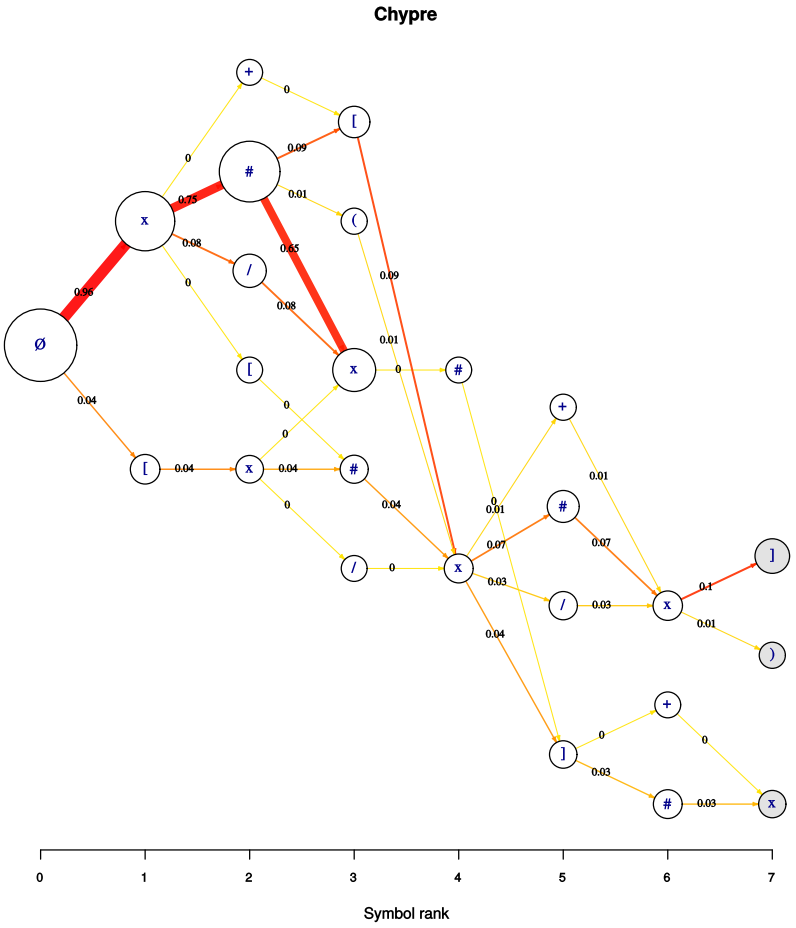

FIGURE 9 – Exemple d'une sortie graphique représentant le diagramme d'un ensemble de séquences.

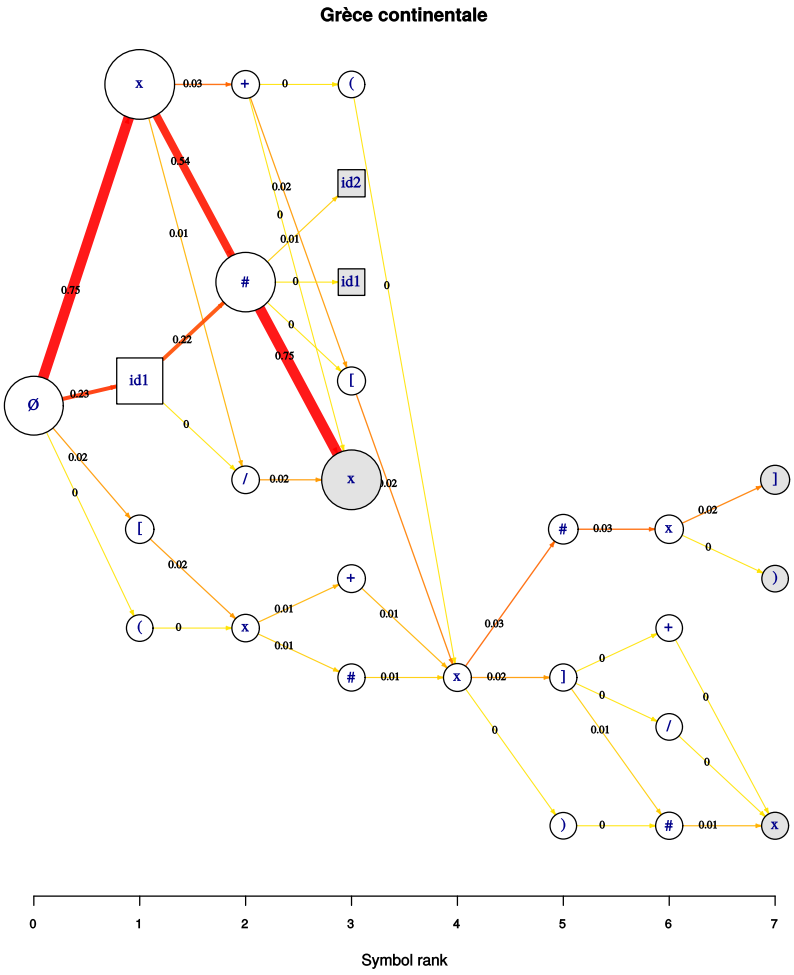

FIGURE 10 – Exemple d'une sortie graphique représentant le diagramme d'un ensemble de séquences avec marquage de deux éléments dont les identifiants ont été saisis dans les champs « Hightlight elements ».

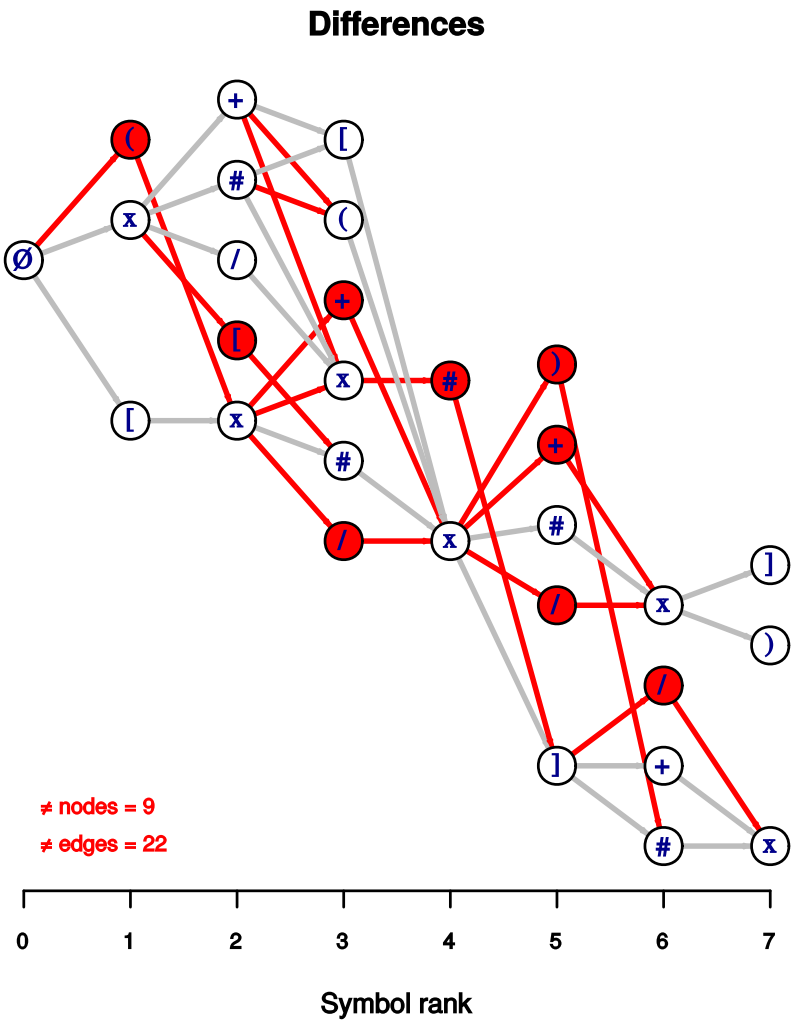

FIGURE 11 – Exemple d'un graphique représentant le diagramme des différences entre deux ensembles de séquences.

rares, mais potentiellement intéressants du point de vue historique. Il est aussi possible d'effectuer des comparaisons entre les séquences relatives à différents espaces, différentes périodes, ou à tout autre ensemble de formules que l'on définirait à partir d'un critère jugé pertinent historiquement (les formules onomastiques relatives à un dieu, à une thématique, *etc.*).

4.2 Diagramme de comparaison

Le diagramme de comparaison facilitent la comparaison visuelle de deux diagrammes de séquences (Figure 11). Étant donnés deux diagrammes de séquences, il permet de mettre en évidence les différences concernant les sommets et les arêtes. Les sommets et arêtes de couleur rouge n'existent que dans l'un des deux diagrammes de séquences comparées ; au contraire, les sommets dont le fond et blanc, et les arêtes grises sont communes aux deux diagrammes de séquence. En bas à gauche du graphique, un décompte des arêtes et sommets différents est affiché.

5 Interprétation des résultats numériques

5.1 Statistiques descriptives individuelles des diagrammes

Une série de statistiques descriptives sont calculées sur les diagrammes. Ces valeurs permettent, d'une part, de caractériser un diagramme donné, et d'autre part, d'identifier objectivement les différences entre plusieurs diagrammes.

5.1.1 Propriétés basiques

Ces valeurs sont calculées quel que soit le paramétrage de l'analyse.

nombre total de séquences : le nombre de séquences total pour la modalité de la variable considérée.

nombre de séquences sélectionnées : le nombre de séquences prises en compte pour construire le diagramme, compte tenu du filtrage par les variables et les longueurs minimales et maximales des séquences données par l'utilisateur. Pour les comparaisons, cet indice permet d'évaluer si les ensembles comparés ont un ordre de grandeur similaire ou différent.

nombre de séquences différentes : le nombre de séquences différentes parmi celles prises en compte pour construire le diagramme. Cette mesure informe à propos de la richesse de l'ensemble de séquences considérées.

longueur maximale des séquences : la longueur, en nombre de symboles, de la plus longue séquence parmi celles prises en compte pour construire le diagramme.

5.1.2 Propriétés du premier élément mis en évidence

Ces valeurs ne sont calculées que si l'identifiant d'un élément a été saisi par l'utilisateur dans le premier champ situé sous « *Highlight elements* ».

Id 1 : nombre de positions : le nombre de position occupées par l'élément considéré. Si, dans l'ensemble de séquences étudiées, il occupe par exemple les deuxième et quatrième, position, alors cette valeur sera de 2.

Id 1 : médiane des degrés entrant : cette mesure est une mesure de la centralité (et donc de l'importance) de l'élément dans l'ensemble de séquence considéré. La valeur renvoyée correspond à la médiane du degré pondéré entrant de l'ensemble des positions occupées par l'élément. Dans un graphe, le degré pondéré entrant correspond aux valeurs des arêtes dirigées vers le sommet considéré.

5.1.3 Richesse globale des types de séquence

Ces valeurs sont calculées seulement si aucun identifiant d'élément n'a été saisi par l'utilisateur dans le premier champ situé sous « *Highlight elements* ».

nombre de sommets : le nombre de sommets contenus dans le graphe, qui correspondent ici à des associations entre symboles et positions dans la séquence. (Le nombre maximal est théoriquement limité : pour 7 positions et 9 symboles différents, le nombre théorique maximal de sommets est de 63. Ce nombre s'abaisse si l'on tient compte de la grammaire d'emploi des symboles.). Cette mesure informe à propos de la richesse de l'ensemble de séquences considérées.

nombre de symboles différents : le nombre de symboles présents parmi les 9 possibles. Cette mesure informe à propos de la richesse de l'ensemble de séquences considérées.

5.1.4 Indices de la spécificité des types de séquences

Ces valeurs sont calculées seulement si aucun identifiant d'élément n'a été saisi par l'utilisateur dans le premier champ situé sous « *Highlight elements* ».

nombre de symboles terminaux : le nombre de symboles apparaissant dans le diagramme en position terminale (les feuilles de l'arbre, en termes de graphes). Par définition, les symboles à la droite du diagramme sont terminaux, puisqu'ils correspondent à la position maximale admise en fonction de la valeur maximale entrée par l'utilisateur pour le paramètre « *Seq. length min/max* ». Toutefois, si l'ensemble de séquences étudié contient des séquences plus courtes, et qui ne correspondent pas au segment initial d'un autre type de séquences plus long, alors d'autres symboles terminaux seront présents dans la diagramme. Ainsi, le nombre symboles terminaux signale la présence de ce type de séquences spécifiques.

nombre de points d'articulation : certains types de séquences peuvent avoir une partie initiale et une partie finale identique, ne différant que par leur partie intermédiaire. Au contraire, des ensembles de séquences peuvent aussi présenter des parties initiales similaires, puis à partir d'une certaine position (que l'on peut se figurer comme une « bifurcation »), des suites complètement différentes. Pour capturer la tendance à se trouver dans l'une ou l'autre de ces situations, le nombre de points d'articulations du graphe est calculé. Les points d'articulations sont les sommets qui, s'ils sont supprimés, divisent le graphe en deux sous-graphes disjoints. N.B. : le symbole initial est supprimé avant le calcul car cette mesure porte sur les sommets et le sommet initial n'existe pas dans la séquence observée dans les sources historiques.

centralisation de degré : un ensemble de séquences peut présenter comporter des symboles positionnés qui se retrouvent dans de nombreuses séquences, et constituent ainsi des « points de passage obligés » ; au contraire, d'autres ensembles de séquences pourraient posséder peu de symboles positionnés de cette sorte, ce qui se traduirait par un diagramme aux cheminements linéaires, possédant moins d'alternatives à chaque position. Afin de mesurer la tendance d'un diagramme à correspondre au premier ou au deuxième cas de figure, la centralisation de degré est calculée. Dans un graphe, la centralisation indique la tendance générale du graphe à être organisé autour de certains sommets centraux. La centralisation peut être calculée de différentes manières [9]. Nous employons la plus simple, basée sur le nombre d'arêtes attachées aux différents sommets (leur degré). *N.B.* : la mesure du degré ne tient pas compte de la pondération des arêtes. En outre, le symbole initial virtuel est supprimé avant le calcul puisque cette mesure porte sur les sommets et que, par définition, toutes les séquences « passent » par le symbole vide initial.

5.1.5 Probabilités

Ces valeurs sont calculées seulement si aucun identifiant d'élément n'a été saisi par l'utilisateur dans le premier champ situé sous « *Highlight elements* ».

probabilité médiane : la valeur médiane des probabilités de transition. Cette mesure porte sur les probabilités observées des transitions entre symboles positionnés. Elle permet d'apprécier si le nombre d'occurrences des différentes transitions possibles dans l'ensemble considéré se répartit plutôt de manière homogène ou si l'ensemble est dominé par certaines transitions. Plus la valeur est haute, plus les syntaxes de l'ensemble de séquences considéré a tendance à être dominé par certaines transitions très fréquentes.

5.2 Distance entre diagrammes

Comme évoqué précédemment, la mesure de la distance entre deux graphes suppose le choix d'une méthode parmi les nombreuses méthodes existantes, qui sera déterminant pour les résultats obtenus. Les résultats numériques donnés à cet égard par *MAP Syntax diagrams* ne doivent donc pas être considérés de la même manière que ceux résultant des mesures descriptives réalisés sur les graphes : si les premiers sont déterminés par des méthodes relativement simples et faisant consensus dans les communautés de spécialistes de l'analyse de graphe, les mesures de distance doivent être considérées comme très dépendantes de la méthode employée pour l'obtenir. Ainsi, un résultat ne pourra être repris et commenté qu'en étant accompagné de la mention de la méthode l'ayant généré. Ces résultats, en dépit de leur dépendance à la méthode employée, peuvent être très utiles pour attirer l'attention sur des différences ou des similarités générales entre des ensembles de séquences ; une fois qu'une relation particulière entre deux ensembles a été identifiée,

9. Voir Freeman 1979.

il est possible de revenir aux statistiques descriptives pour comprendre plus en détail ce qui distingue ces deux ensembles.

6 Conclusion

L'application *MAP Syntax diagrams* constitue un outil relativement puissant et flexible pour l'analyse de la structure d'ensembles de séquences onomastiques telles que définies dans le projet MAP. Les fonctions proposées intègrent un éventail de possibilités allant de statistiques descriptives robustes à des méthodes de comparaison plus sophistiquées, en passant par l'identification des positions syntaxiques d'éléments sémantiques donnés. L'interface, simplement accessible à travers un navigateur web, rend ces fonctions aisément accessibles et facilite l'exploration rapide des données. Les possibilités de comparaison ne sont limitées que par l'imagination des utilisateurs, qui entreront dans l'application les tableaux de variables pertinentes pour des questions de recherche données. On dispose ainsi d'un instrument rigoureux pour l'étude de la variabilité observable dans la manière de construire les titulatures divines en différents temps et lieux de l'Antiquité.

Références

Chang, Winston, Joe Cheng, J.J. Allaire, Yihui Xie et Jonathan McPherson [2019], *shiny : Web Application Framework for R*, R package version 1.4.0, https://CRAN.R-project.org/package=shiny.

Csárdi, Gábor et Tamás Nepusz [2006], « The igraph Software Package for Complex Network Research », *InterJournal*, 1695, 5, p. 1-9, http://igraph.org.

Freeman, Linton C. [1979], « Centrality in Social Networks I: Conceptual Clarification », *Social Networks*, 1, p. 215-239, DOI : 10.1016/0378-8733(78)90021-7.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas Müller et Matthias Studer [2011], « Analyzing and Visualizing State Sequences in R with TraMineR », *Journal of Statistical Software*, 40, 4, p. 1-37.

Hammond, David K., Yaniv Gur et Chris R. Johnson [2013], « Graph Diffusion Distance: A Difference Measure for Weighted Graphs Based on the Graph Laplacian Exponential Kernel », dans *Proceedings of the IEEE global conference on information and signal processing (GlobalSIP'13)*, Austin (Tex.), p. 419-422, DOI : 10.1109/GlobalSIP.2013.6736904.

Ipsen, Mads et Alexander S. Mikhailov [2002], « Evolutionary Reconstruction of Networks », *Physical Review E*, 66, 4, p. 039901, DOI : 10.1103/PhysRevE.66.046109.

Jurman, Giuseppe, Roberto Visintainer et Cesare Furlanello [2010], *An Introduction to Spectral Distances in Networks (extended version)*, https://arxiv.org/abs/1005.0103.

Lebreton, Sylvain et Corinne Bonnet [2019], « Mettre les polythéismes en formules ? À propos de la base de données Mapping Ancient Polytheisms », *Kernos*, 32, p. 267-296, DOI : 10.4000/kernos.3163.

Martínez, Johann H. et Mario Chavez [2019], « Comparing complex networks: in defence of the simple », *New Journal of Physics*, 21, p. 013033, DOI : 10.1088/1367-2630/ab0065.

Murtagh, Fionn [1985], *Multidimensional Clustering Algorithms*, Compstat Lectures, Vienna : Physika Verlag.

You, Kisung [2019], *NetworkDistance : Distance Measures for Networks*, R package version 0.3.2, https://CRAN.R-project.org/package=NetworkDistance.