Chaînes de référence et structuration textuelle

quelques indicateurs...

Bruno Oberlé

Journée d'Étude Democrat

14 juin 2019

UNIVERSITÉ DE STRASBOURG

democrat

Introduction

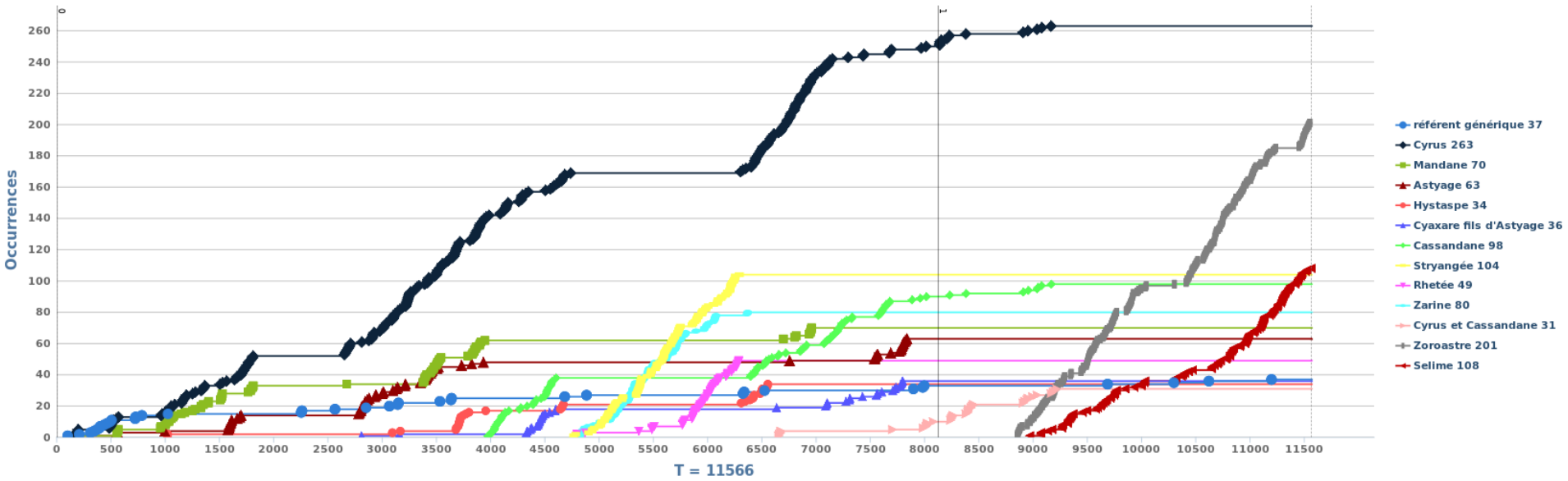
Contexte

- **étude de la relation entre chaînes de référence et paragraphes**
- **corpus**
  - 5 textes issus du corpus Democrat
  - du 12e au 20e siècle
  - Énéas, Jehan de Paris, Pantagruel, Princesse de Clèves, Voyage de Cyrus, Ventre de Paris et Némoville

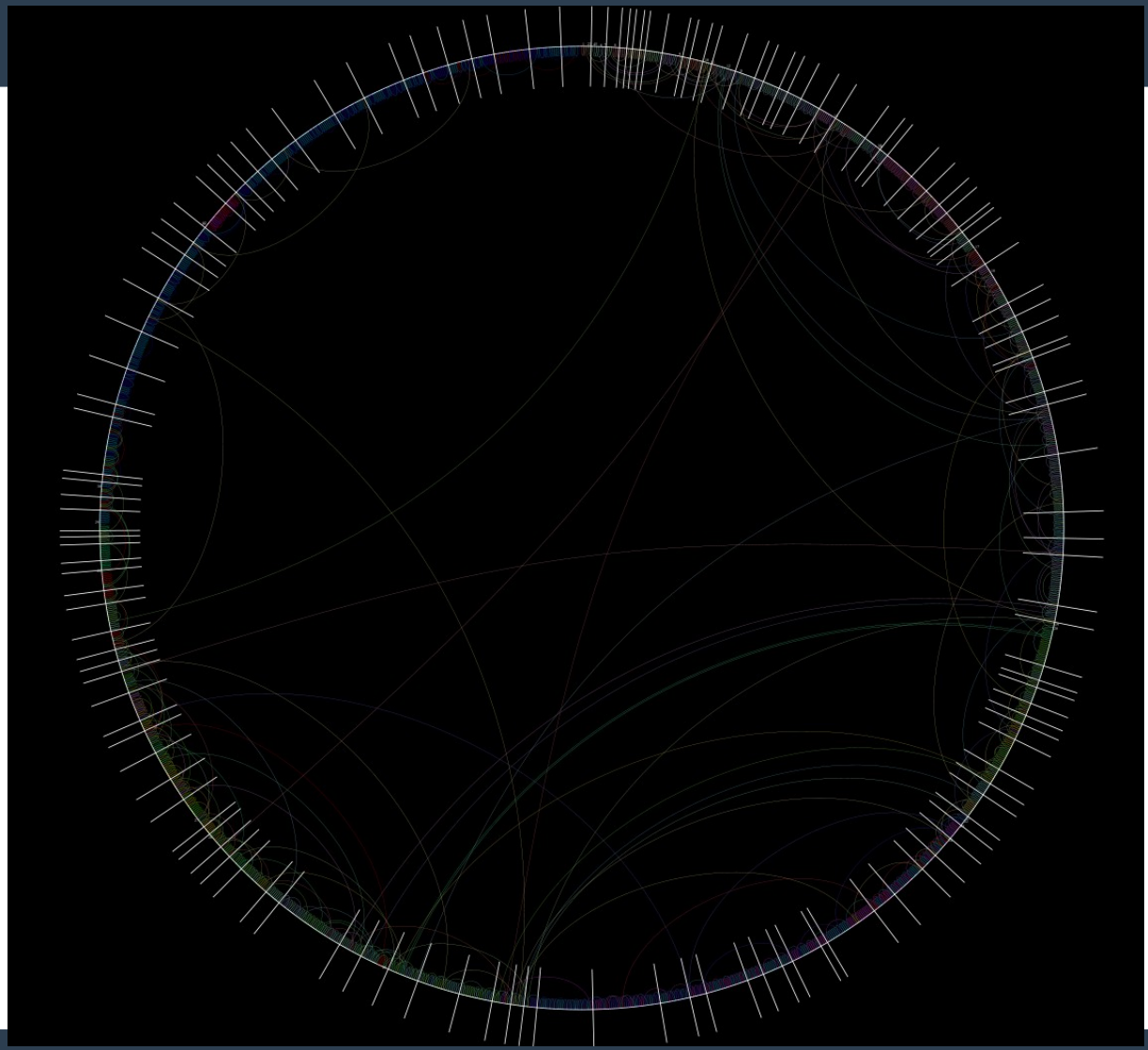
Observation

- **essai de visualisation des chaînes de référence**
  - “progression” (TXM)
  - “en cercle” *(chord diagram)*
  - “carte”
- **question: Visuellement, quelle est la relation entre chaînes et paragraphes?**

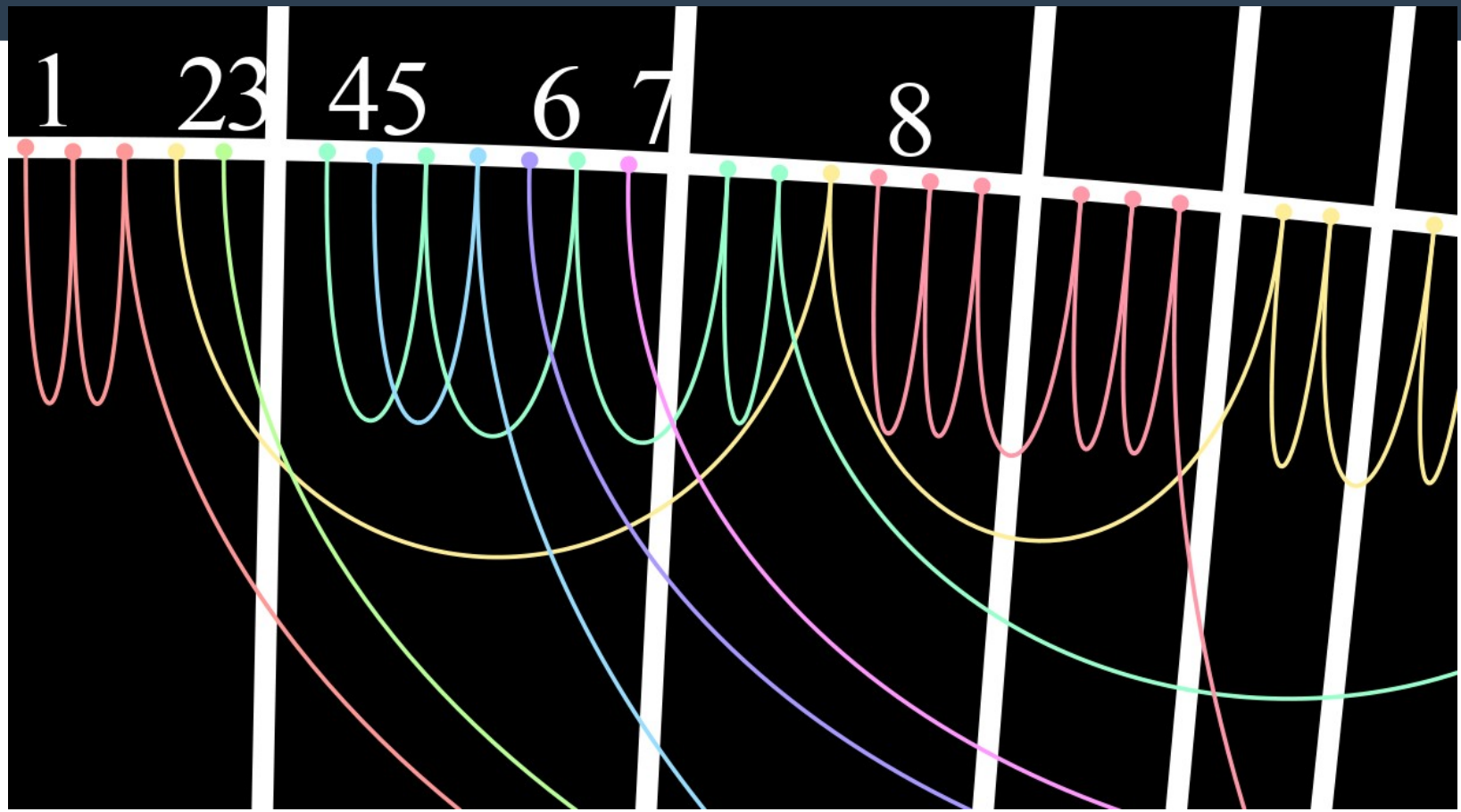
Observation: Progression



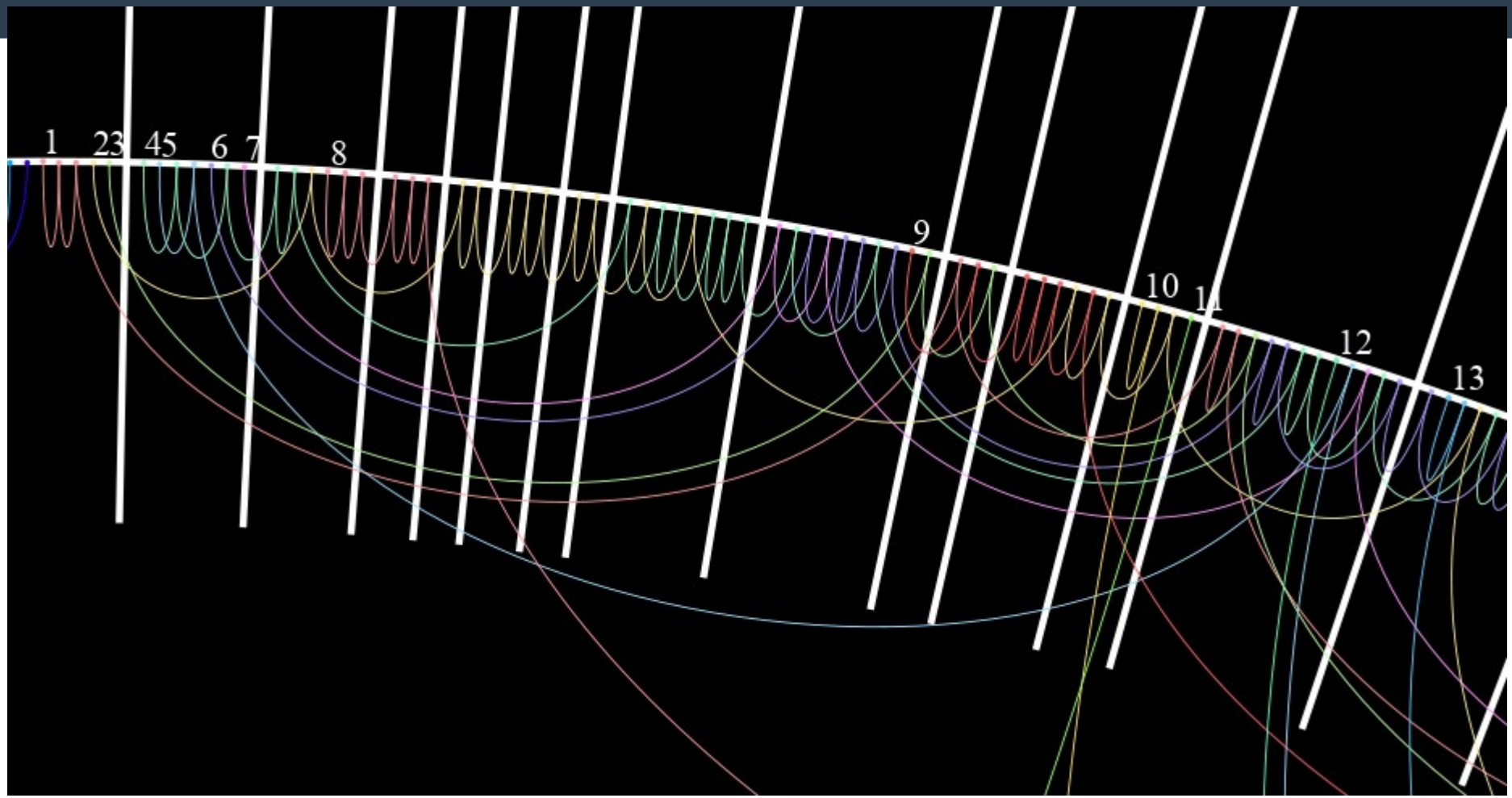
Observation: chord



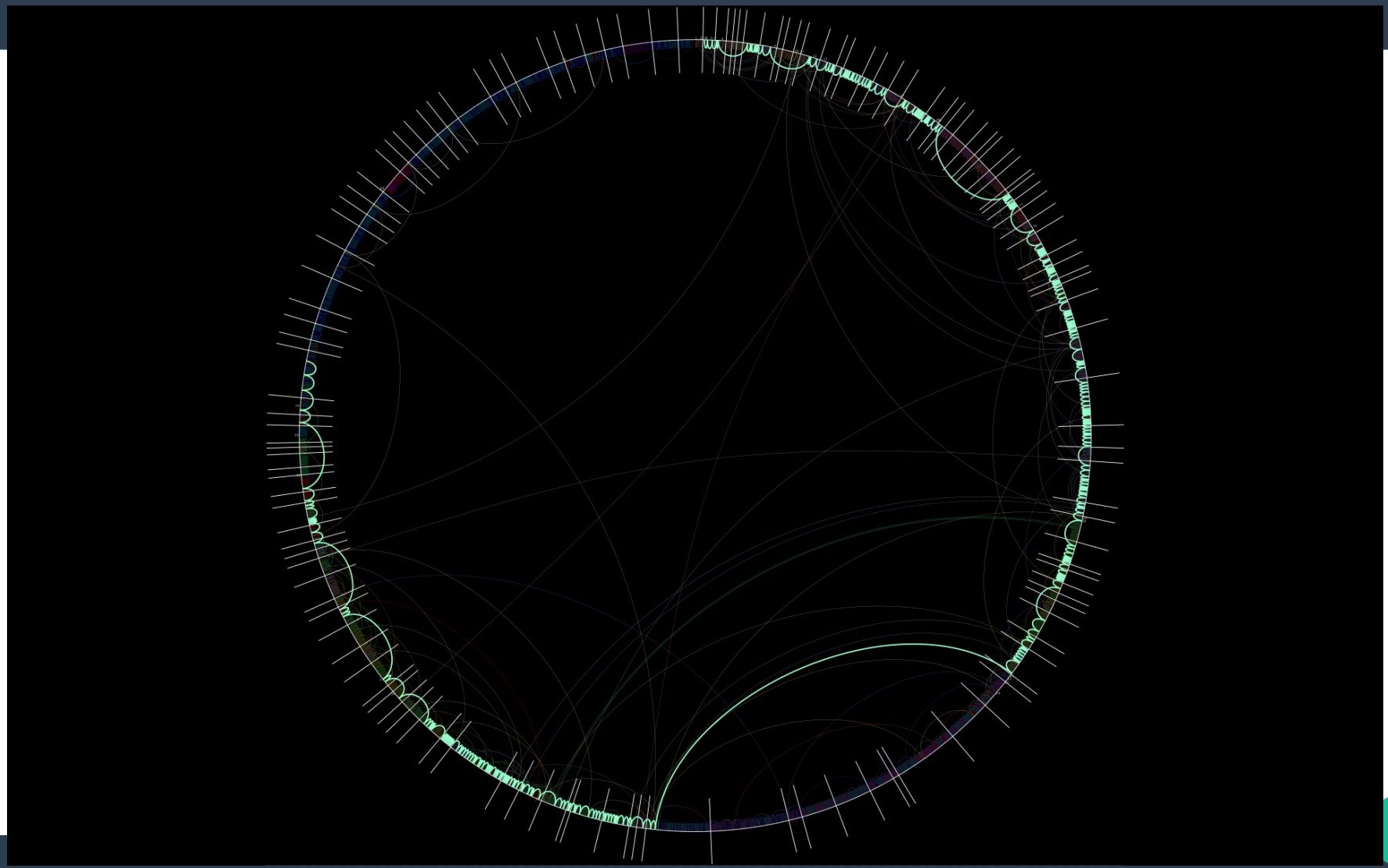
Observation: chord



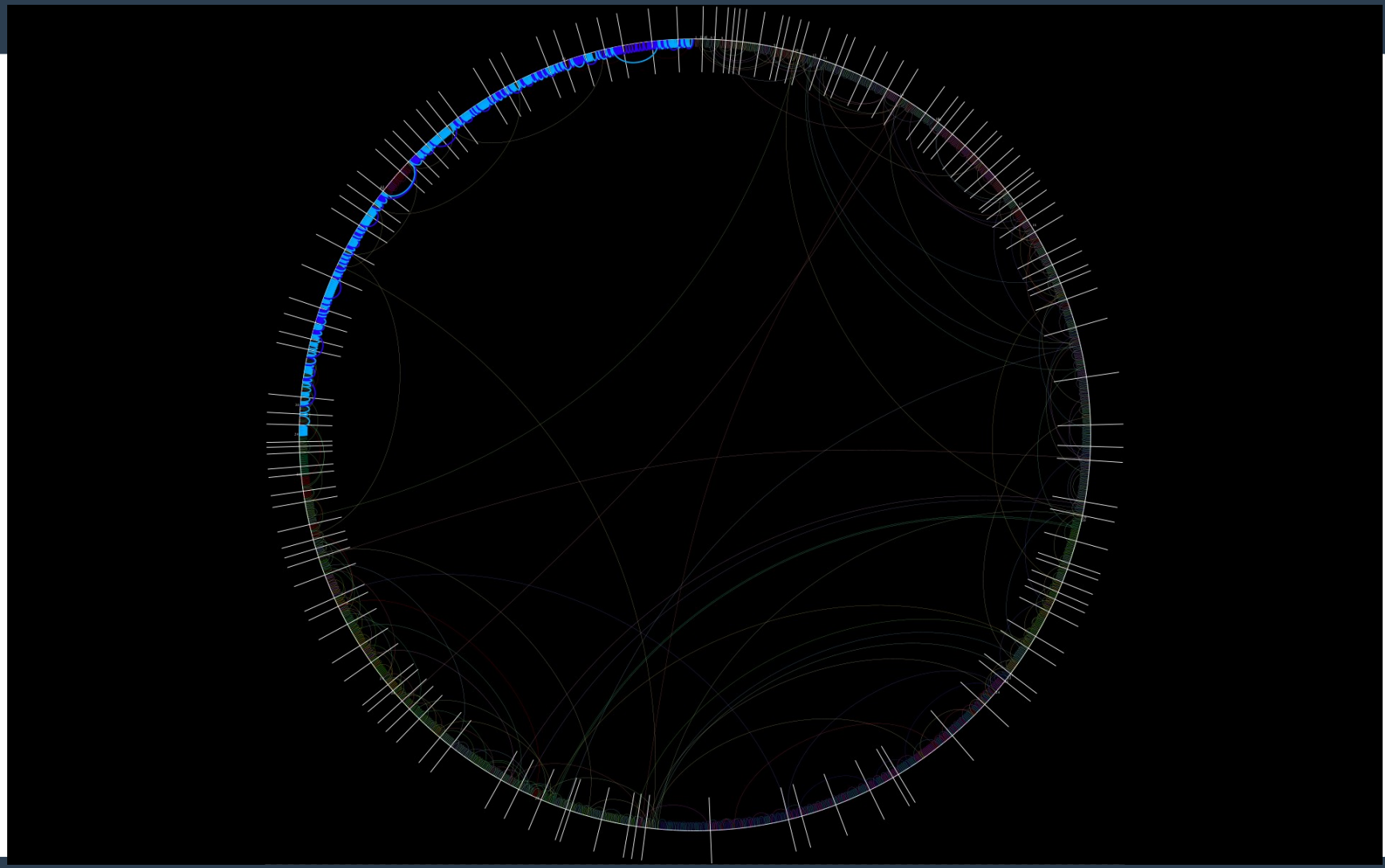
Observation: chord



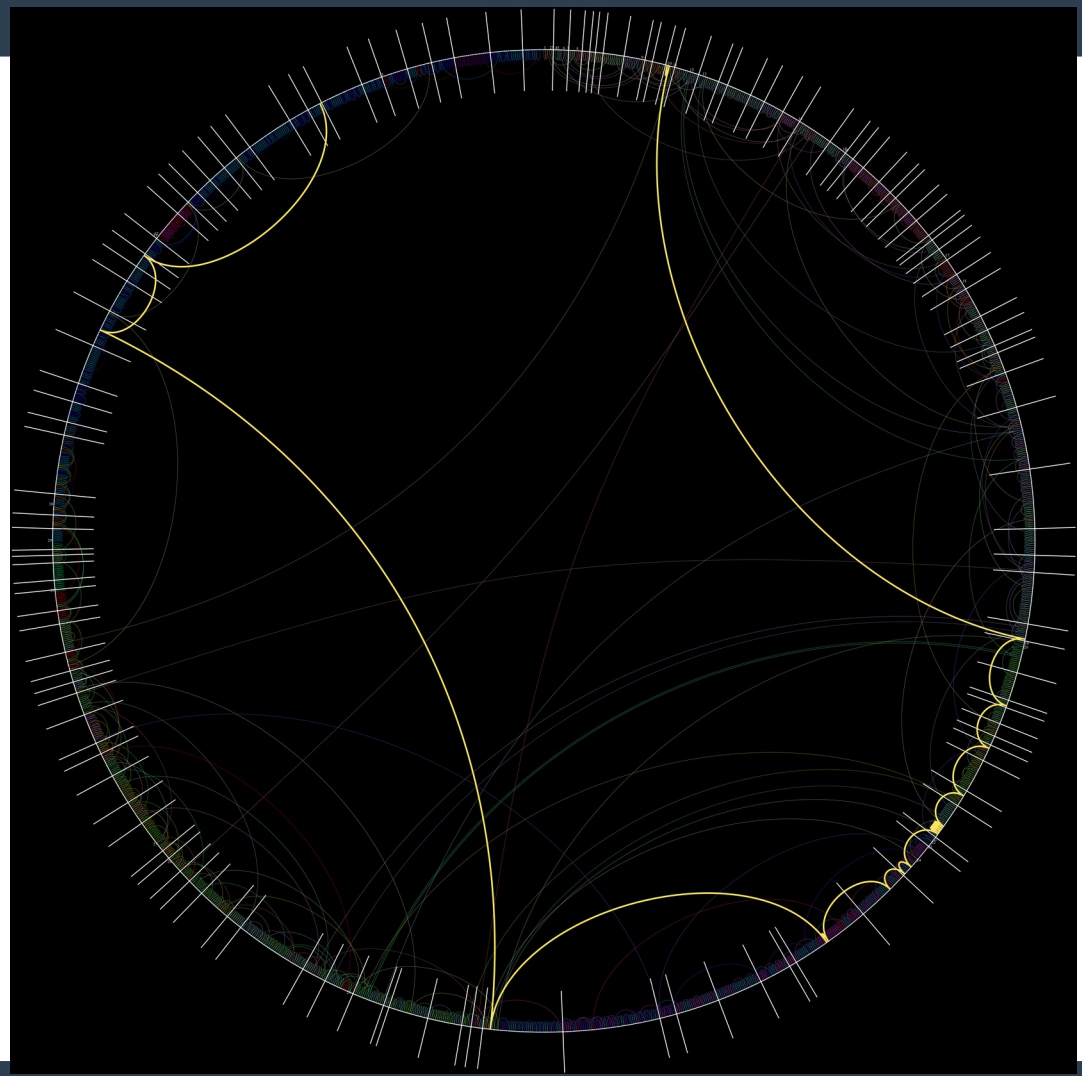
Observation: chord



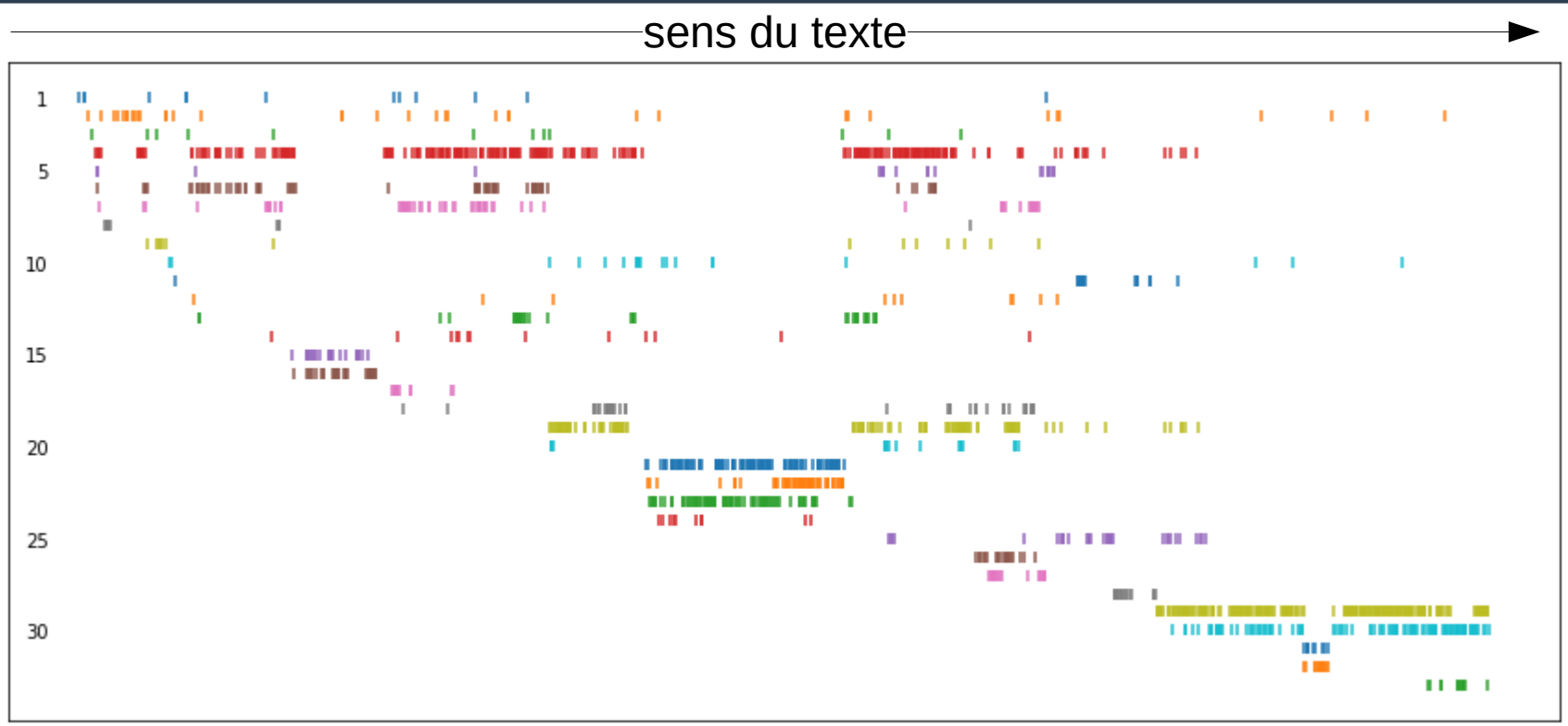
Observation: chord



Observation: chord

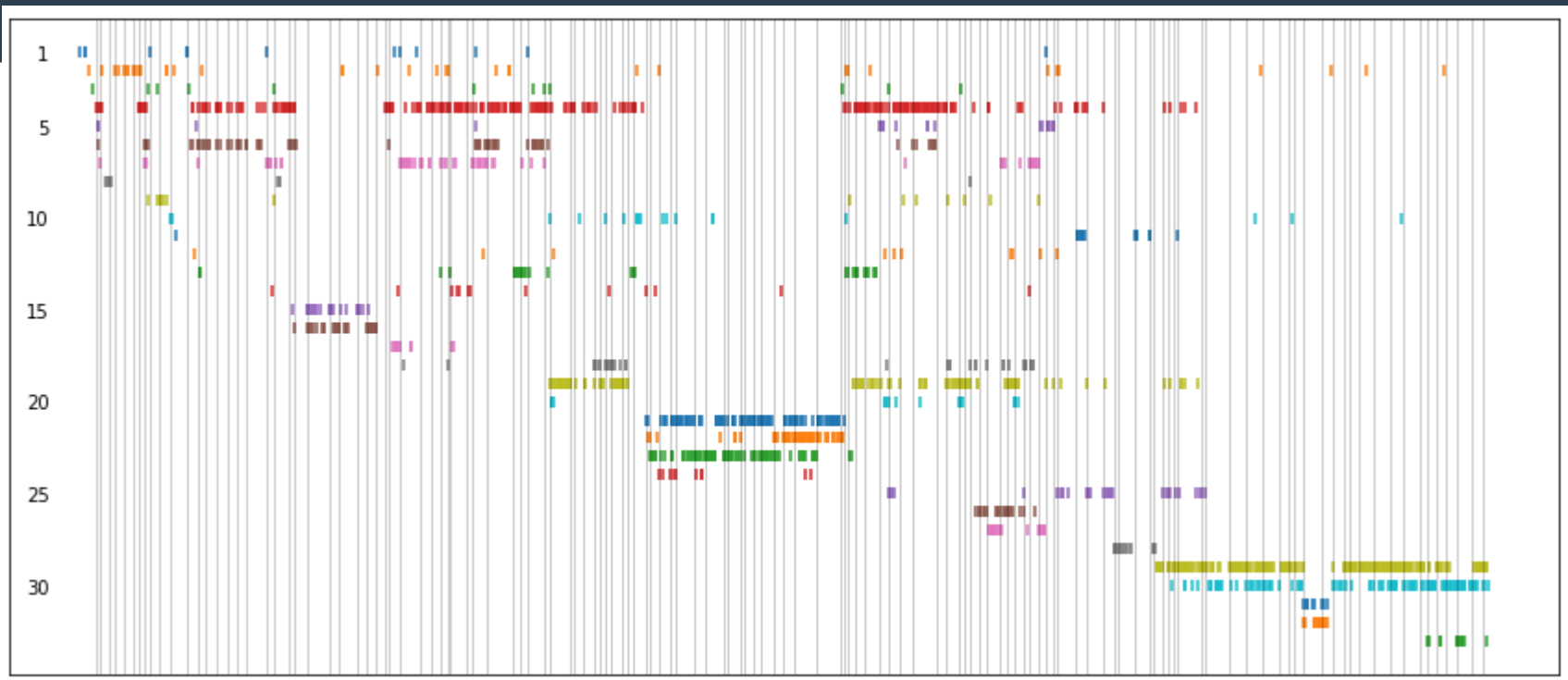


Observation: carte



(1) la_Medie, (2) referent_generique, (3) Ecbatane, (4) Cyrus, (5) Cambyse, (6) Mandane, (7) Astyage, (8) les_perses, (9) la_cour_d_Ecbatane, (10) l_amour, (11) les_mages, (12) Perse, (13) Hystaspe, (14) les_medes, (15) Logis, (16) Sigee, (17) Merodac, (18) Cyaxare_fils_d_Astyage, (19) Cassandane, (20) Farnaspe, (21) Stryangee, (22) Rhetee, (23) Zarine, (24) Zarine_et_Styangee, (25) Cyrus_et_Cassandane, (26) Araspe, (27) Harpage, (28) philosophes, (29) Zoroastre, (30) Selime, (31) les_lyciens, (32) les_lyciennes, (33) les_lyciens_generiques_,

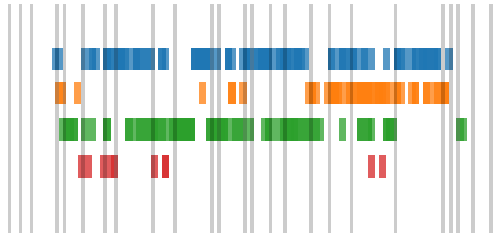
Observation: carte



(1) la_Medie, (2) referent_generique, (3) Ecbatane, (4) Cyrus, (5) Cambyse, (6) Mandane, (7) Astyage, (8) les_perses, (9) la_cour_d_Ecbatane, (10) l_amour, (11) les_mages, (12) Perse, (13) Hystaspe, (14) les_medes, (15) Logis, (16) Sigee, (17) Merodac, (18) Cyaxare_fils_d_Astyage, (19) Cassandane, (20) Farnaspe, (21) Stryangee, (22) Rhetee, (23) Zarine, (24) Zarine_et_Styangee, (25) Cyrus_et_Cassandane, (26) Araspe, (27) Harpage, (28) philosophes, (29) Zoroastre, (30) Selime, (31) les_lyciens, (32) les_lyciennes, (33) les_lyciens_generiques_,

Constats

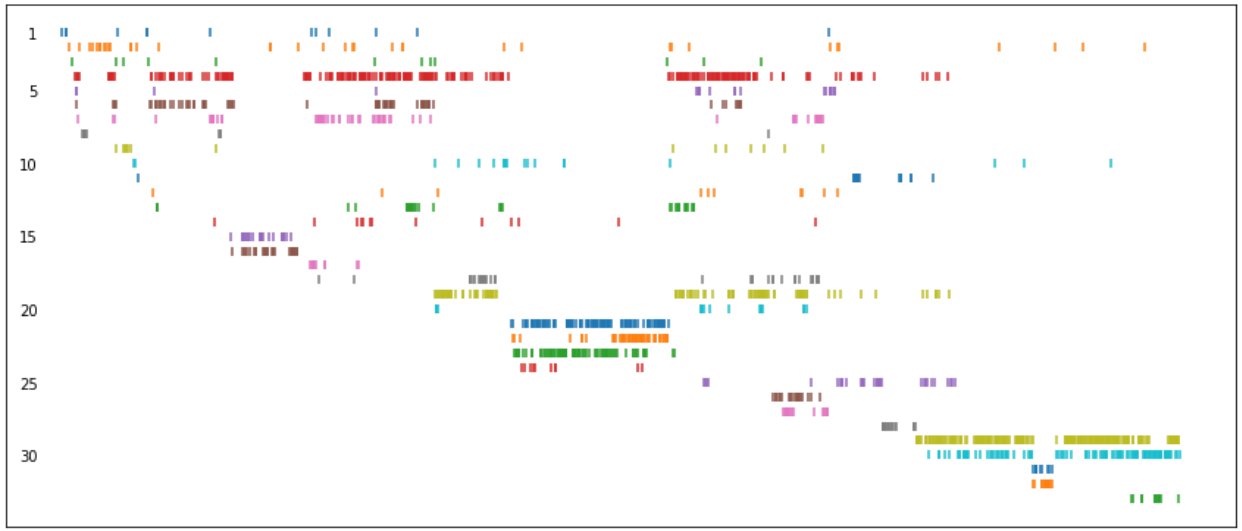
- certaines chaînes semblent former des regroupements de mentions (séquences)



- ces regroupements traversent plusieurs paragraphes
- une structure textuelle entre par et chapitre, non marquée typographiquement?

Objectifs

- **peut-on segmenter le texte à partir des chaînes de référence?**
- **comment caractériser ces segments?**



Intérêts

- **relation entre chaînes et structure textuelle:**
  - ne correspondent pas aux paragraphes
  - ne correspondent pas aux chapitres
  - niveau intermédiaire
- **une étape de l’étude des chaînes de référence dans les paragraphes:**
  - paragraphes de transition
  - paragraphes de rupture
- **utile pour les textes sans paragraphe:**
  - textes anciens
  - oral

Buts de la communication

- **présentation de quelques indicateurs des chaînes:**
  - pour segmenter le texte
  - pour étudier les paragraphes
- **illustration sur un texte (Voyage de Cyrus)**

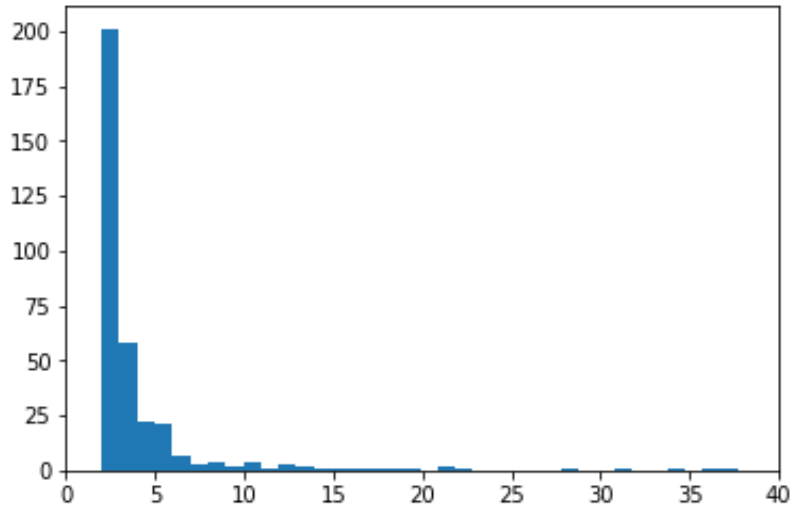
Outline

- **présentation du texte d’étude**
- **essai de segmentation:**
  - rafales de Lafon
  - “flyers”
  - changement de chaînes principales
  - sous-chaînes
- **deux autres indicateurs prometteurs**
  - enchevêtrement de Lafon
  - noms vs pronoms
- **marché aux indicateurs**

Texte d’étude

Cyrus de Ramsay

- **“Voyage de Cyrus”, Ramsay, 18e siècle**
- **extrait des 11 570 premiers tokens**
- **annoté dans le cadre de Democrat**
- **chaînes:**
  - 1 156 singletons
    + 351 chaînes
    = 1 507 référents
  - distribution des chaînes:



Cyrus de Ramsay

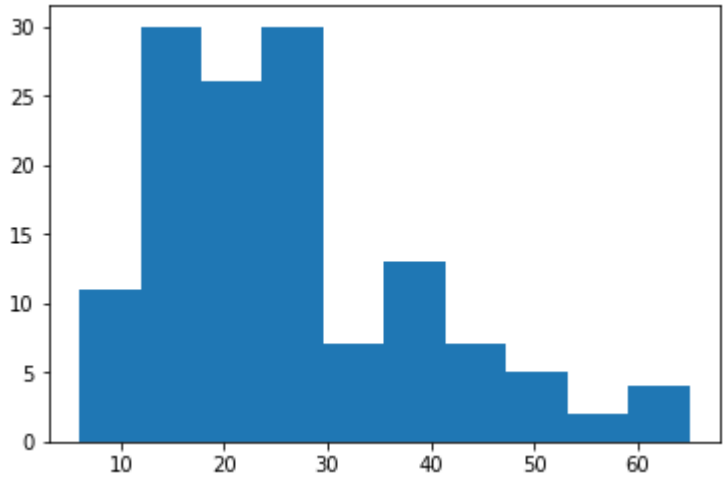
- **chaînes:**
  - longueur moyenne: 13 maillons
  - distance intermaillonnaire moyenne: 157 tokens
- **dans cette présentation:**

uniquement
les chaînes de 10 maillons ou plus

(1) la_Medie, (2) referent_generique, (3) Ecbatane, (4) Cyrus, (5) Cambyse, (6) Mandane, (7) Astyage, (8) les_perses, (9) la_cour_d_Ecbatane, (10) l_amour, (11) les_mages, (12) Perse, (13) Hystaspe, (14) les_medes, (15) Logis, (16) Sigee, (17) Merodac, (18) Cyaxare_fils_d_Astyage, (19) Cassandane, (20) Farnaspe, (21) Stryangee, (22) Rhetee, (23) Zarine, (24) Zarine_et_Styangee, (25) Cyrus_et_Cassandane, (26) Araspe, (27) Harpage, (28) philosophes, (29) Zoroastre, (30) Selime, (31) les_lyciens, (32) les_lyciennes, (33) les_lyciens_generiques_,

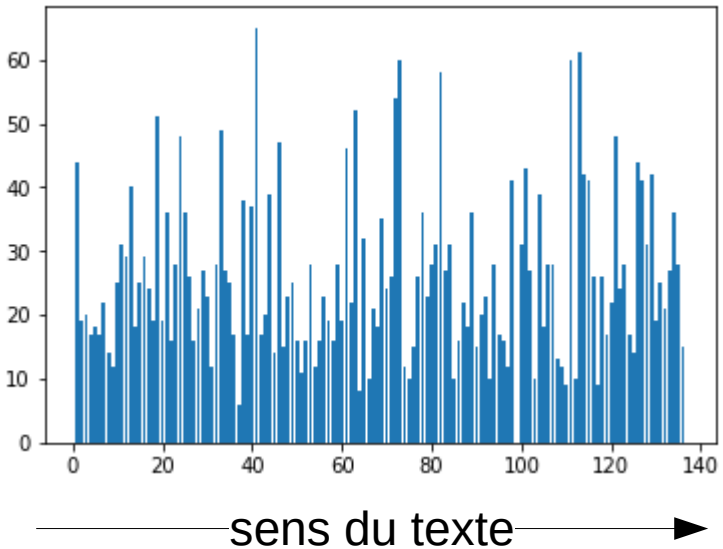
Cyrus de Ramsay

- **paragraphes:**
  - nombre: 137
  - longueur moyenne: 2.83 phrases
  - distribution (mentions)



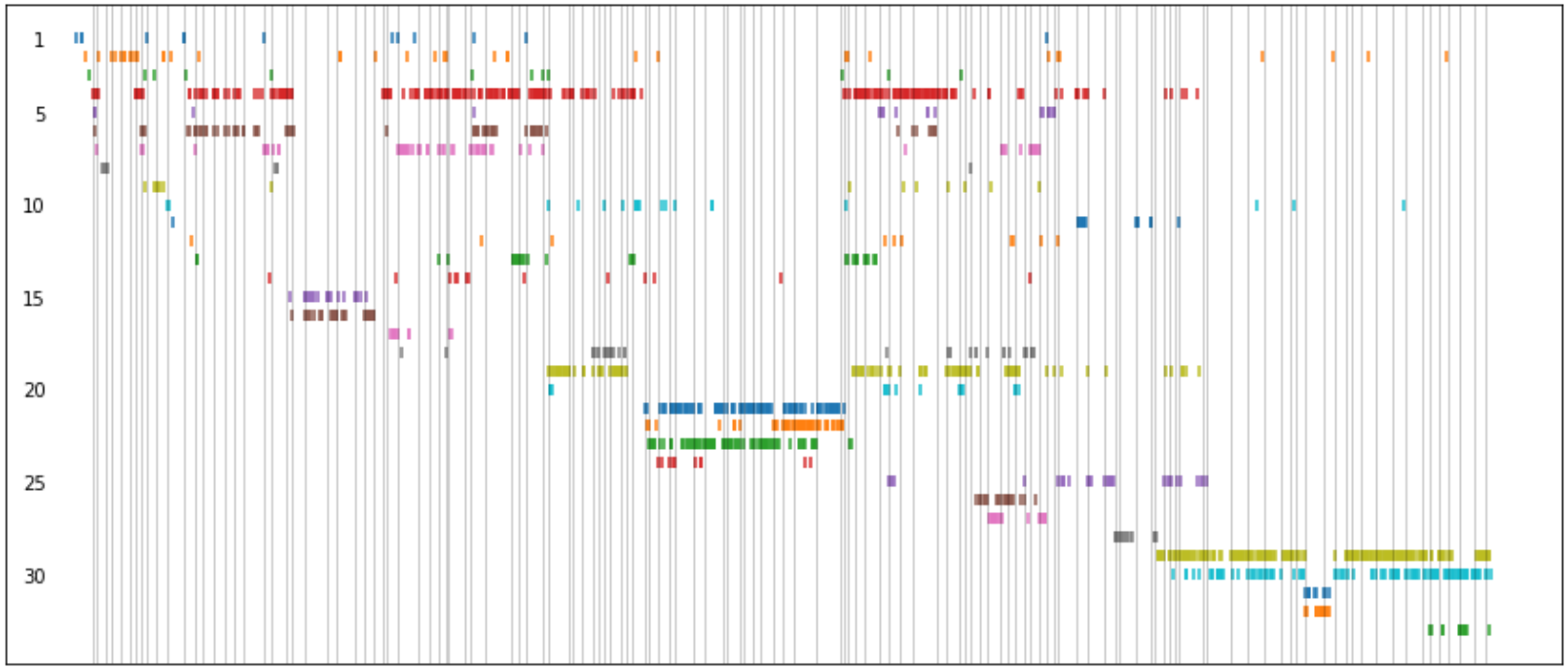
Cyrus de Ramsay

- **des paragraphes inégaux tout au long du texte**
- **nombre de mentions par paragraphe, dans l'ordre du texte:**



Cyrus de Ramsay

- **objectif: au-delà du paragraphe, comme trouver une structure textuelle avec les chaînes de référence?**



Méthodologie

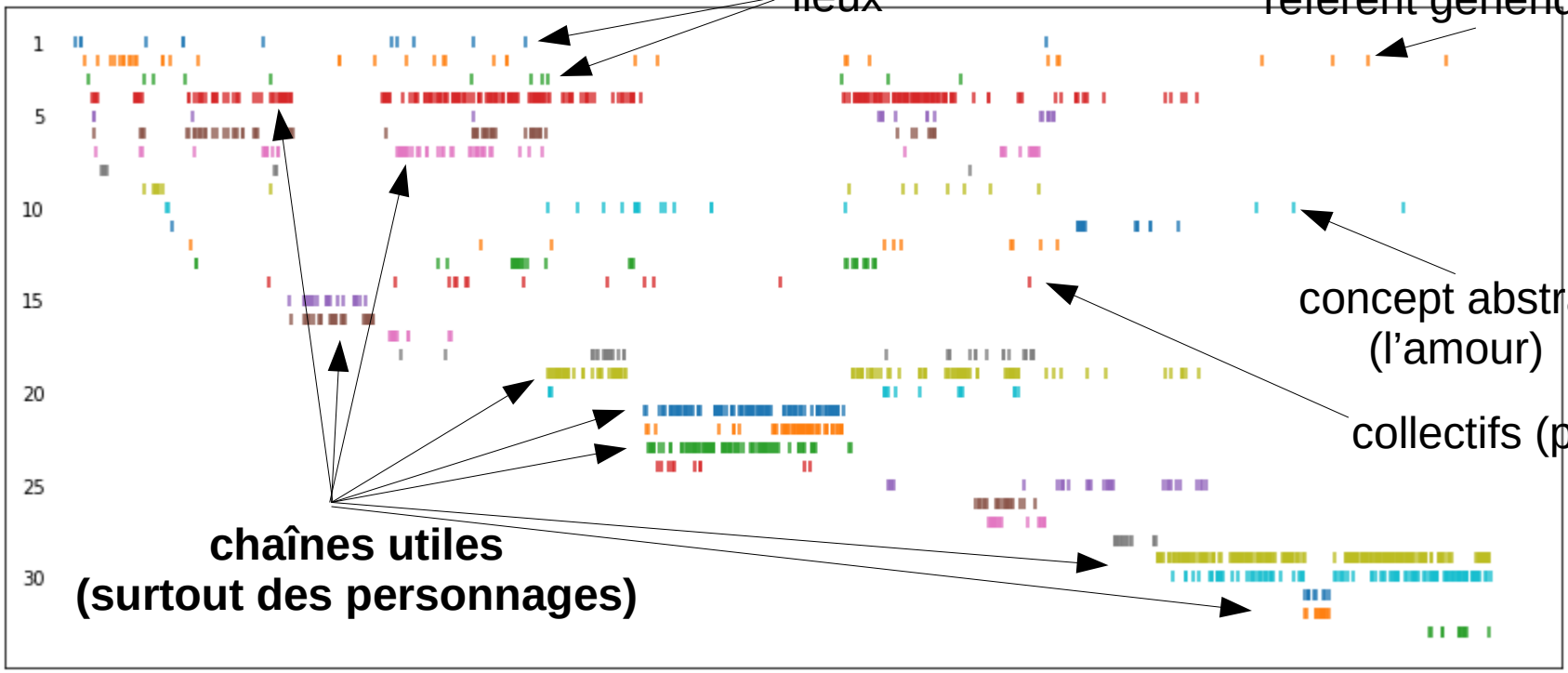
Méthodologie:

1. les rafales de Lafon

Différents types de chaînes

- **toutes les chaînes ne sont pas créées égales**
  - certaines sont espacées
  - d’autres ne sont constituées que de pronoms
  - certaines semblent plus aptes à la segmentation du texte
  - d’autres semblent être un frein à une telle segmentation

Différents types de chaînes



(1) la_Medie, (2) referent_generique, (3) Ecbatane, (4) Cyrus, (5) Cambyse, (6) Mandane, (7) Astyage, (8) les_perses, (9) la_cour_d_Ecbatane, (10) l_amour, (11) les_mages, (12) Perse, (13) Hystaspe, (14) les_medes, (15) Logis, (16) Sigee, (17) Merodac, (18) Cyaxare_fils_d_Astyage, (19) Cassandane, (20) Farnaspe, (21) Stryangee, (22) Rhetee, (23) Zarine, (24) Zarine_et_Styangee, (25) Cyrus_et_Cassandane, (26) Araspe, (27) Harpage, (28) philosophes, (29) Zoroastre, (30) Selime, (31) les_lyciens, (32) les_lyciennes, (33) les_lyciens_generiques_,

Différents types de chaînes

- **toutes les chaînes ne sont pas utiles pour l’étude de la structure textuelle**
- **comment trouver les bonnes?**

Rafales de Lafon

- **Lafon, *Dépouillements et Statistiques en lexicométrie*, 1984**
- **visuellement, on voit bien une différence entre les deux chaînes suivantes:**
  - “régulière” (vert)
  - “en rafale” (rouge)

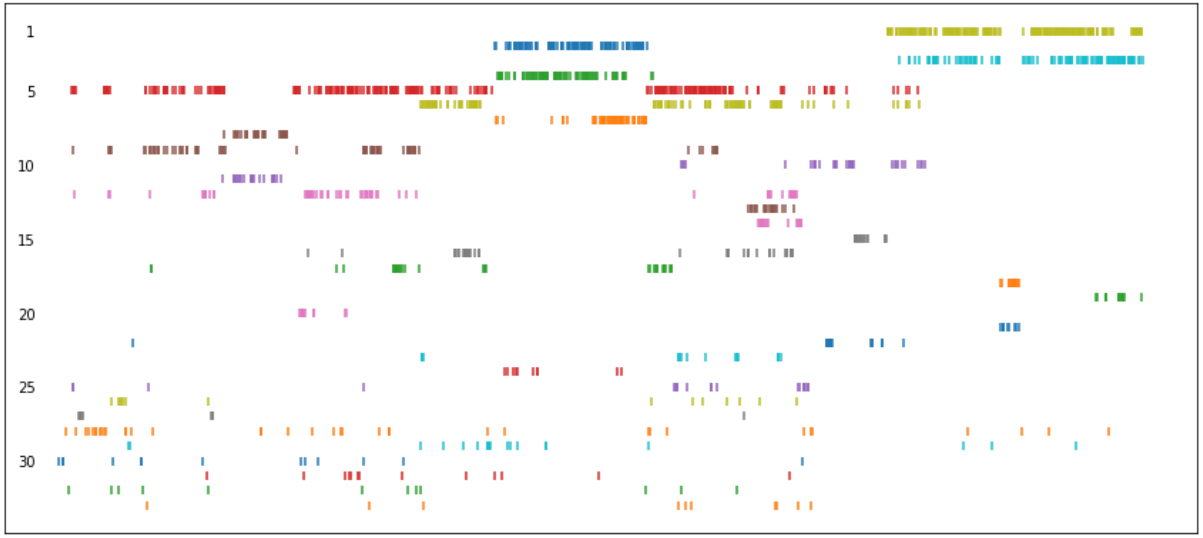
Rafales de Lafon

- **Lafon propose un indicateur pour mesurer la répartition d'éléments dans un texte**
  - indicateur lexicométrique
  - qui a été adapté aux chaînes de référence
- **"rafales de Lafon":**
  - mesure de la probabilité d'avoir une répartition "en rafales" plutôt qu'une répartition régulière (espérée)

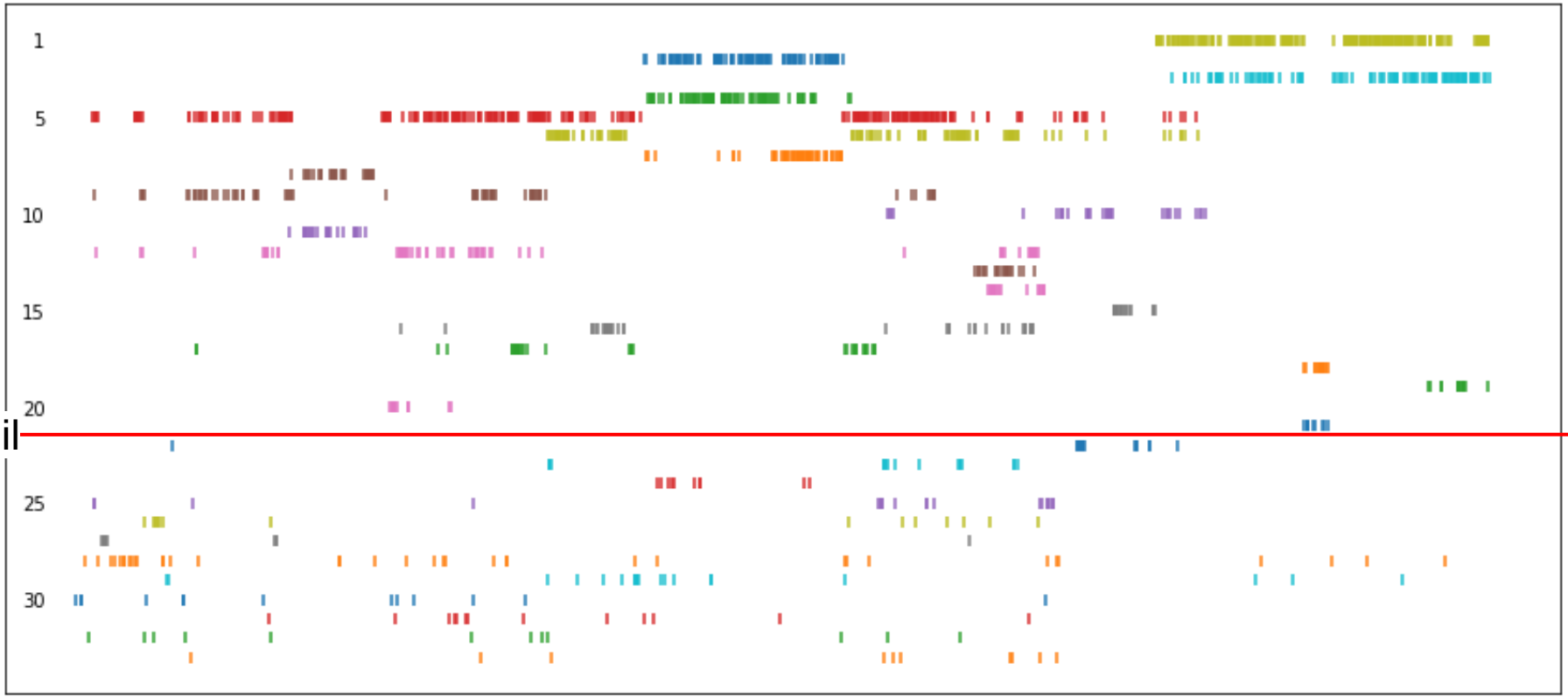
Rafales de Lafon

- **calcul pour les chaînes du texte**
- **tri des chaînes en fonction de leur “rafalité” (les plus rafaleuses en haut)**

  (en grand sur le slide suivant)



Rafales de Lafon



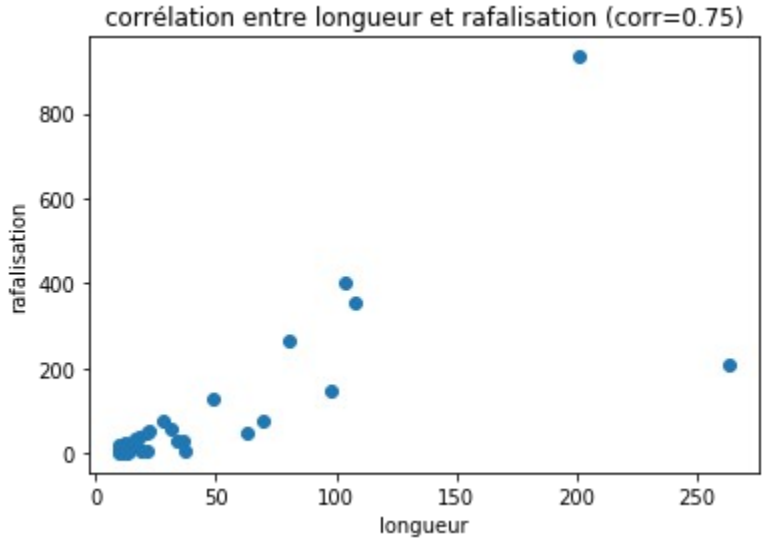
(1) Zoroastre, (2) Stryangee, (3) Selime, (4) Zarine, (5) Cyrus, (6) Cassandane, (7) Rhetee, (8) Sigee, (9) Mandane, (10) Cyrus_et_Cassandane, (11) Logis, (12) Astyage, (13) Araspe, (14) Harpage, (15) philosophes, (16) Cyaxare_fils_d_Astyage, (17) Hystaspe, (18) les_lyciennes, (19) les_lyciens_generiques_, (20) Merodac, (21) les_lyciens, (22) les_mages, (23) Farnaspe, (24) Zarine_et_Styangee, (25) Cambyse, (26) la_cour_d_Ecbatane, (27) les_perses, (28) referent_generique, (29) l_amour, (30) la_Medie, (31) les_medes, (32) Ecbatane, (33) Perse,

Rafales de Lafon

- **constat:**
  - les chaînes les plus rafaleuses semblent définir des segments
  - les référents semblent:
    - ou bien alterner
    - ou bien être parallèles
  - **gardons les 21 chaînes les plus rafaleuses pour notre étude** (seuil défini pour l'instant "visuellement", plus d'études sont nécessaires pour trouver un seuil automatiquement)

Rafales de Lafon

- **limite: trop forte corrélation entre la longueur de la chaîne et la rafalisation?**



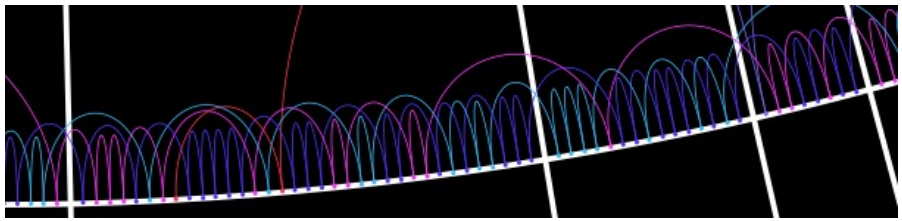
Méthodologie:

2. les flyers

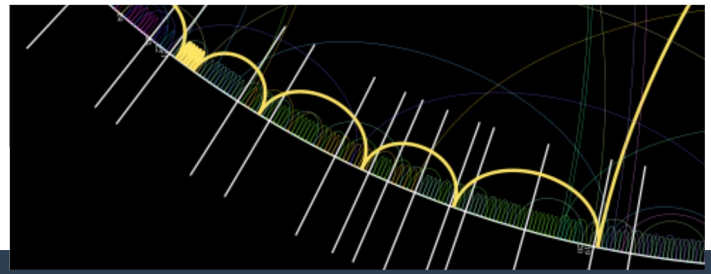
Flyers

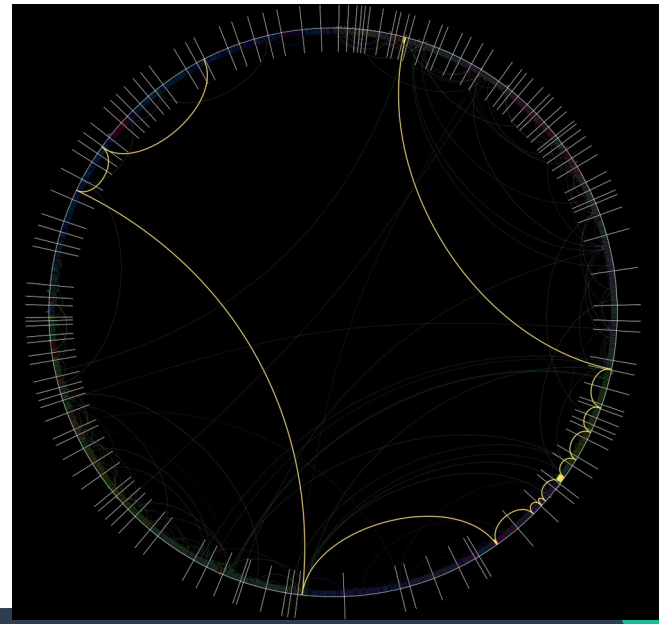
- **les chaînes font des bonds:**
  - quelle est la grandeur de ces bonds?
  - mesure de la distance au dernier maillon

petits bonds



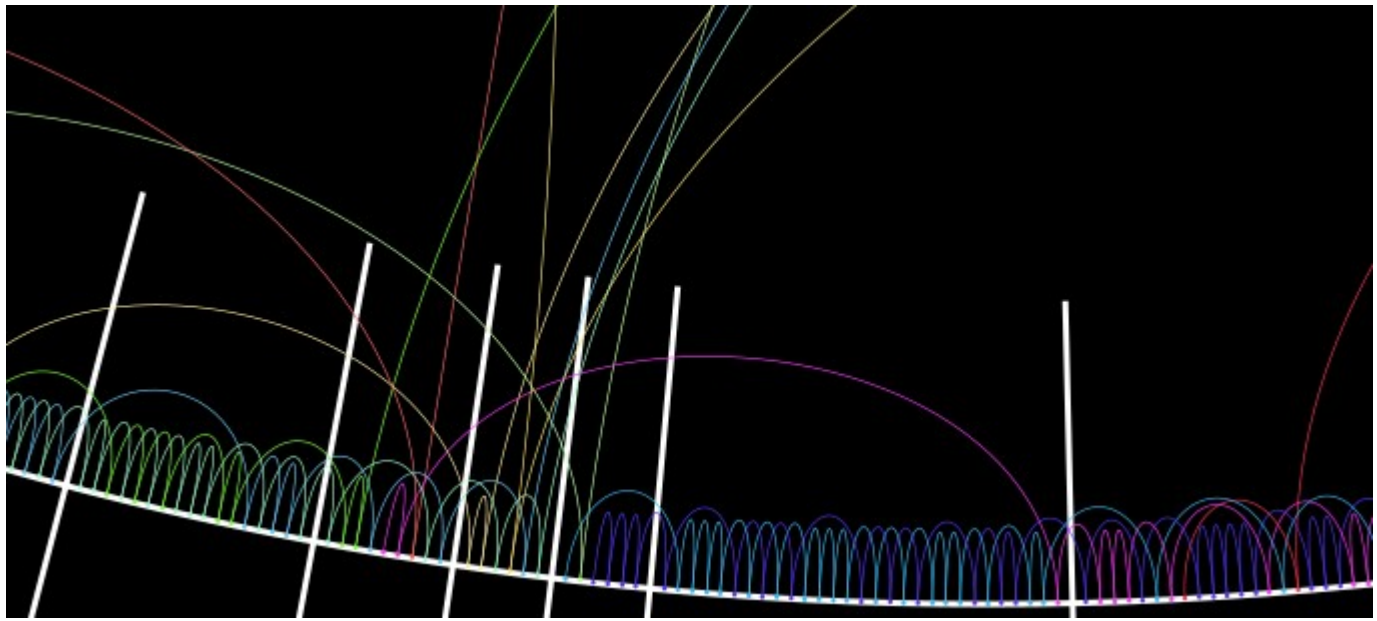
grands bonds





Flyers

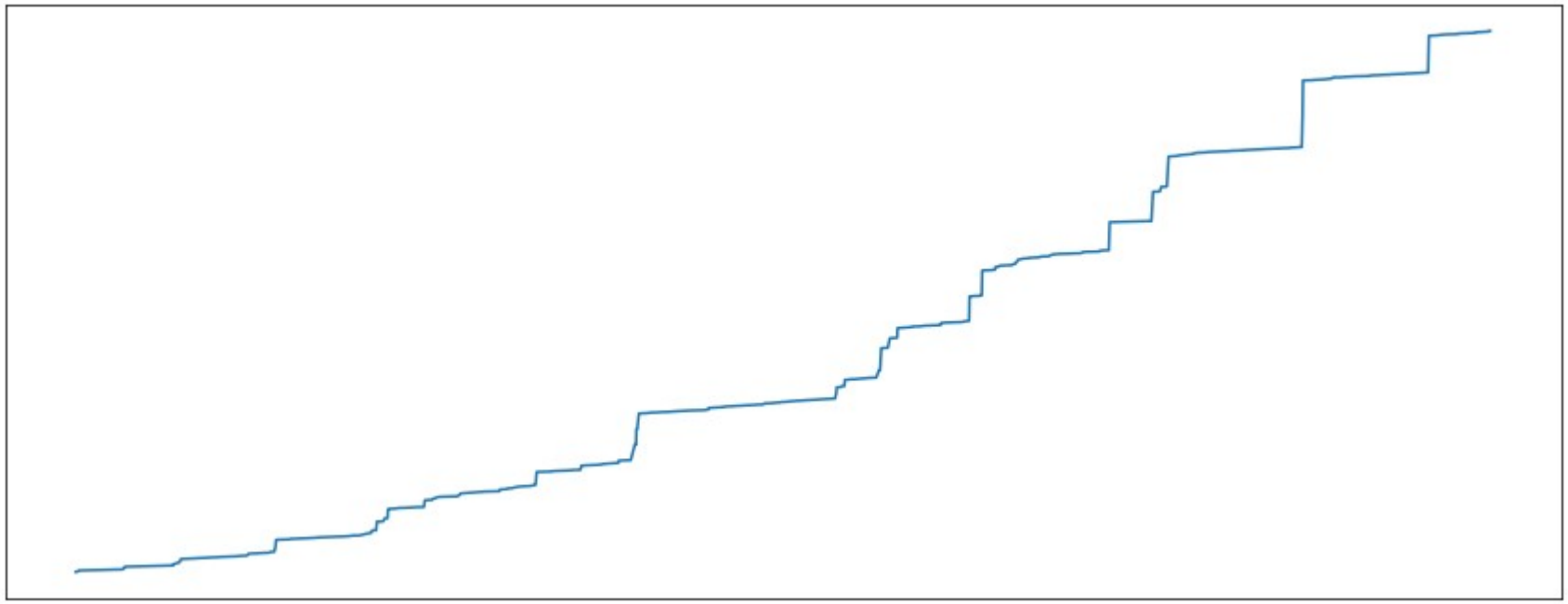
- **hypothèse: une concentration de flyers à longue distance est symptomatique d’une rupture thématique**



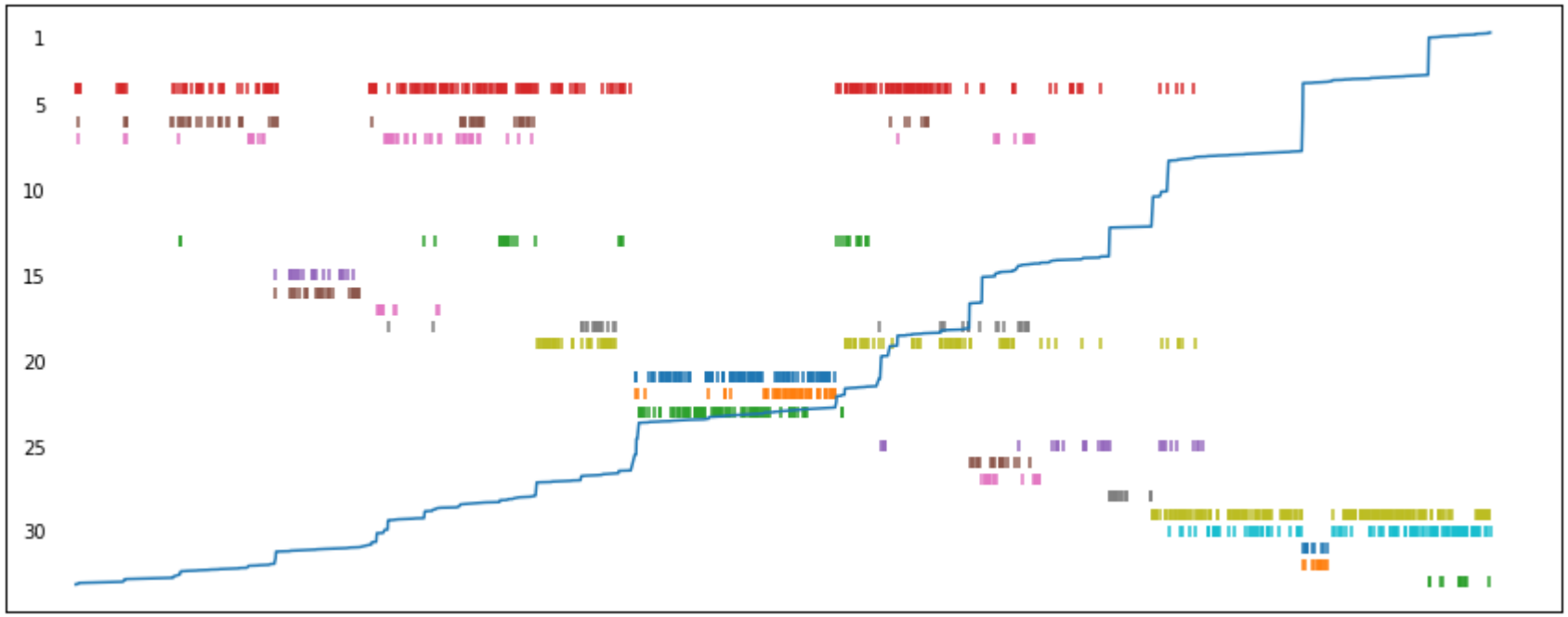
Flyers

- **courbe des flyers:**
  - à chaque nouvelle mention, on ajoute la distance à la dernière mention
  - si la mention n’a pas de précédente mention, on ajoute la distance au début du texte
  - on ne considère que les chaînes rafaleuses
- **TLDR: “plus ça monte brusquement, plus il y a des flyers qui viennent de loin”**
- **constat: les flyers semblent définir des ruptures**

Flyers



Flyers



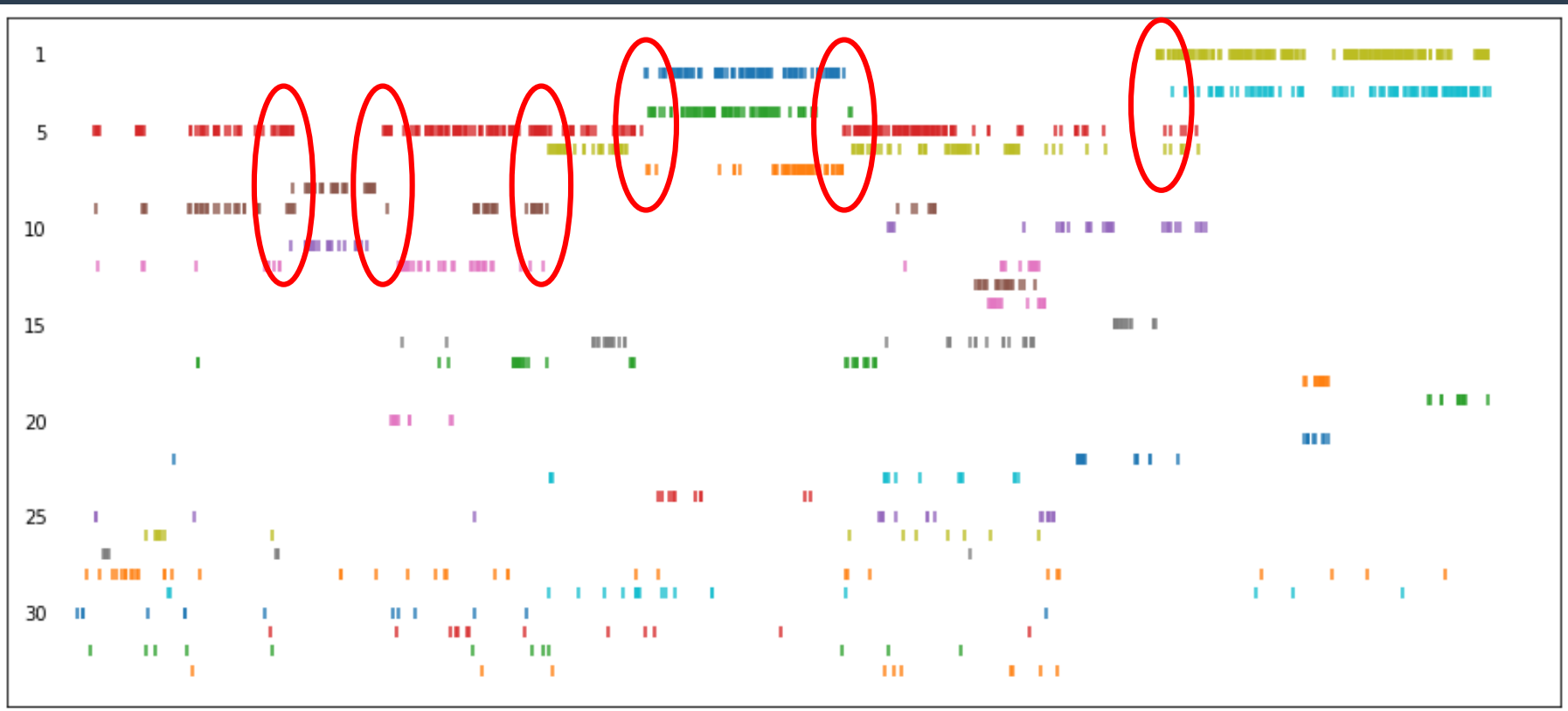
Méthodologie:

3. changement de chaînes principales

Changement de chaînes principales

- **constat:**
  - alternance de référents
  - comme au théâtre: il y a changement de scène lorsqu’il y a changement (entrée, sortie) de personnages

Changement de chaînes principales

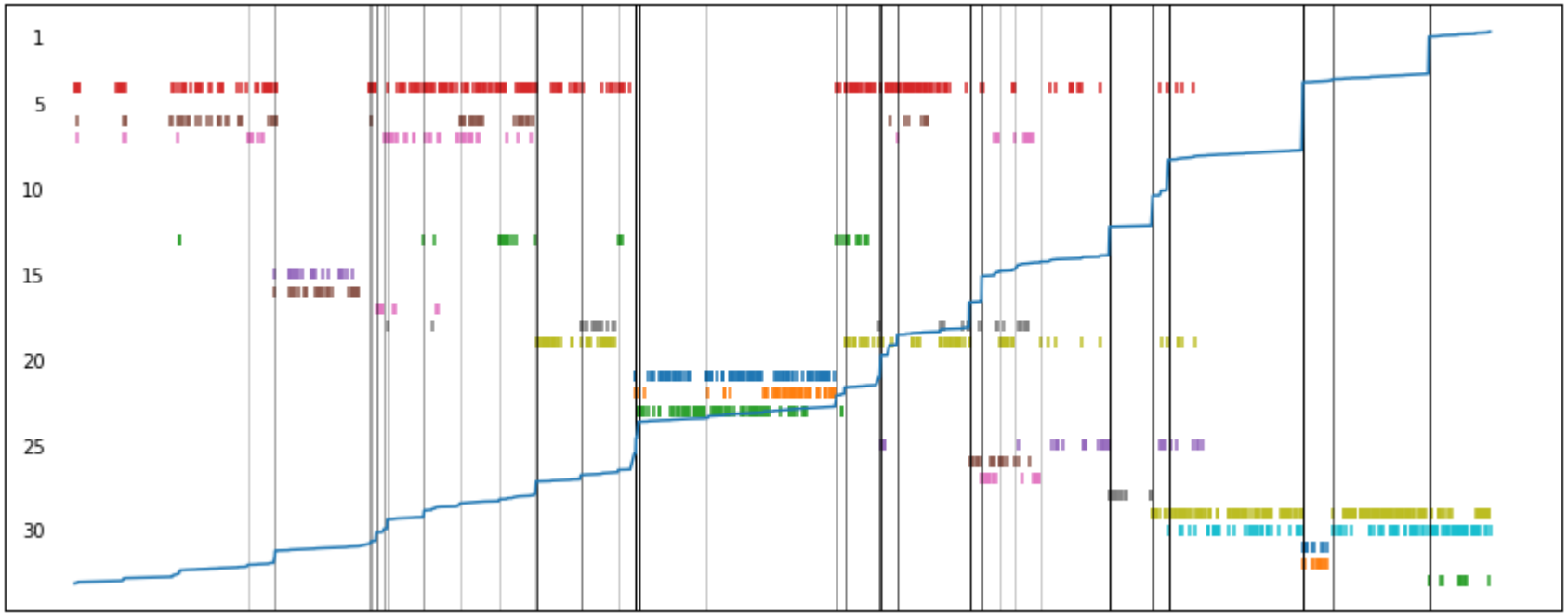


(1) Zoroastre, (2) Stryangee, (3) Selime, (4) Zarine, (5) Cyrus, (6) Cassandane, (7) Rhetee, (8) Sigee, (9) Mandane, (10) Cyrus_et_Cassandane, (11) Logis, (12) Astyage, (13) Araspe, (14) Harpage, (15) philosophes, (16) Cyaxare_fils_d_Astyage, (17) Hystaspe, (18) les_lyciennes, (19) les_lyciens_generiques_, (20) Merodac, (21) les_lyciens, (22) les_mages, (23) Farnaspe, (24) Zarine_et_Styangee, (25) Cambyse, (26) la_cour_d_Ecbatane, (27) les_perses, (28) referent_generique, (29) l_amour, (30) la_Medie, (31) les_medes, (32) Ecbatane, (33) Perse,

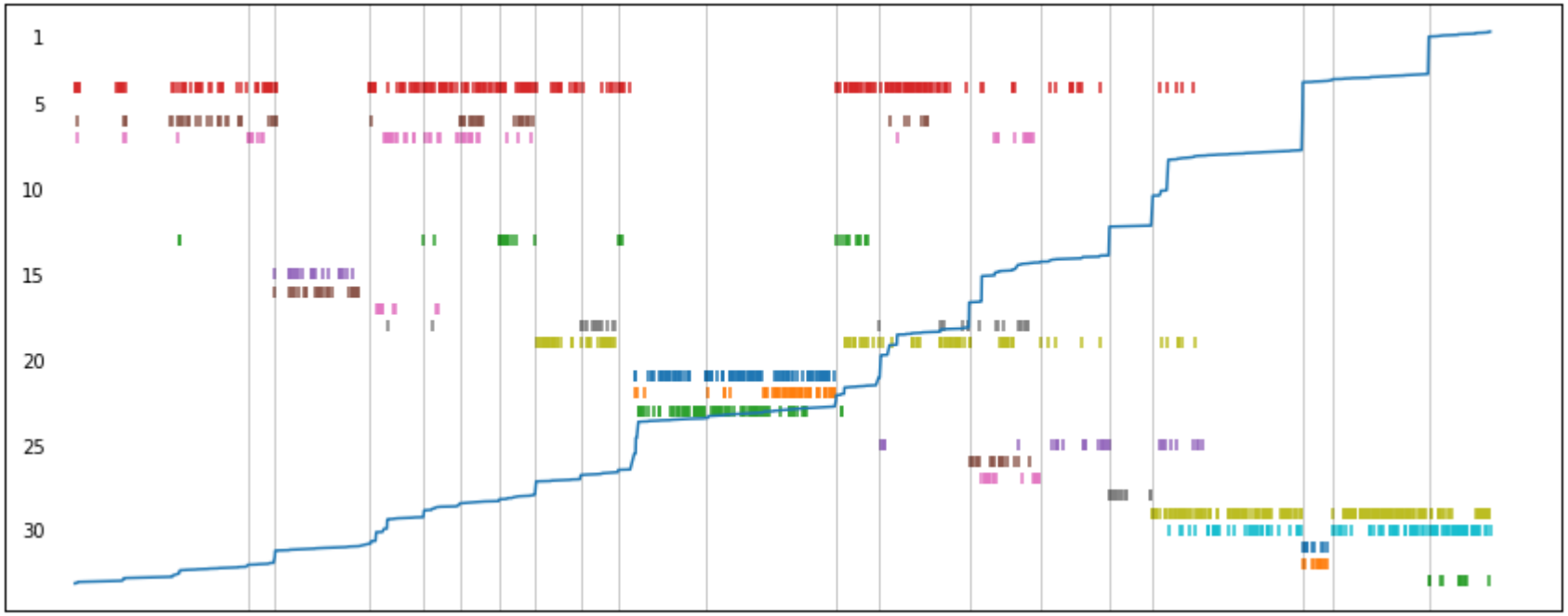
Changement de chaînes principales

- **découverte des points d’échange:**
  - en s’appuyant sur les chaînes rafaleuses
  - par couple de chaînes (en prenant les chaînes deux à deux)
  - à l’aide de n-grams
- **constat: correspondance entre flyers et alternance de référents**

Changement de chaînes principales



Changement de chaînes principales



Méthodologie:

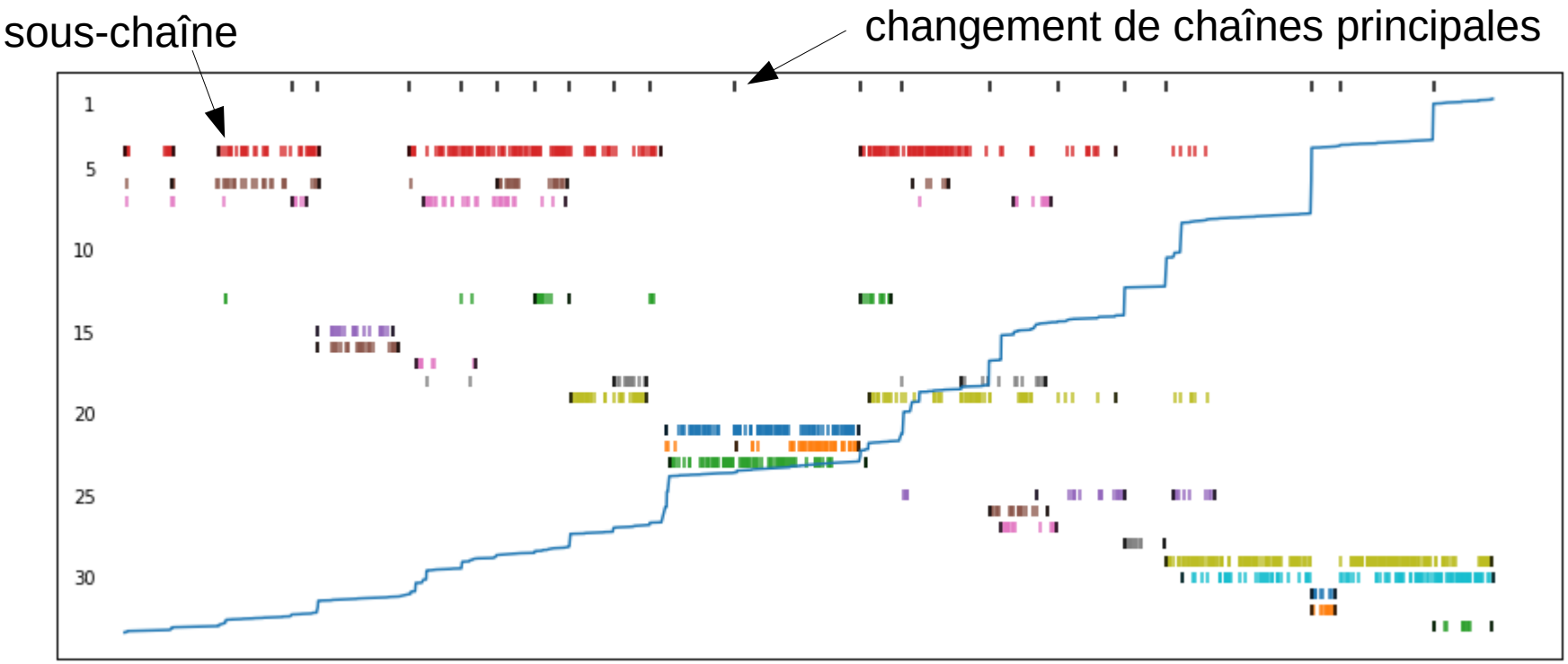
4. sous-chaînes

Sous-chaînes

- **les chaînes rafaleuses peuvent être segmentées en sous-chaînes**
- **objectif: utiliser les frontières de sous-chaînes pour identifier des ruptures thématiques**
- **méthode:**
  - de petites distances entre les mentions d'une même rafale
  - de grandes distances entre les rafales
  - un algorithme de clustering pour identifier la limite entre petites et grandes distances

Sous-chaînes

- **résultat (les frontières de sous-chaînes sont en noir)**



Sous-chaînes

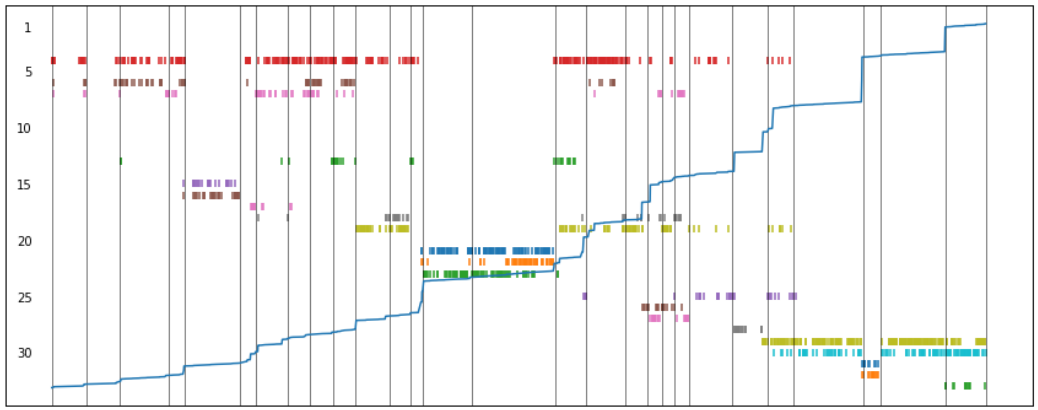
- **constat: correspondance avec les changements de chaînes principales**

Résultats

(segmentation)

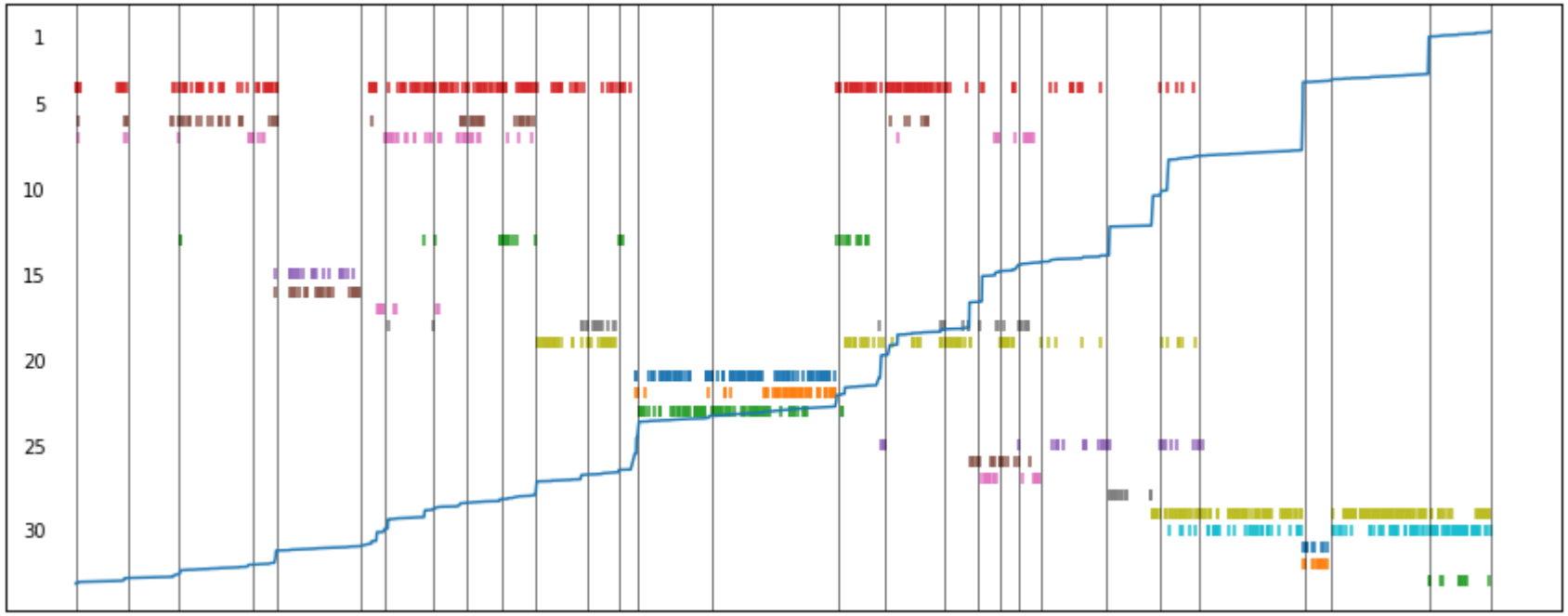
Segmentation du texte

- **calcul des segments à l'aide:**
  - des changements de chaînes principales
  - des frontières de sous-chaînes
- **alignements sur les frontières de phrases**

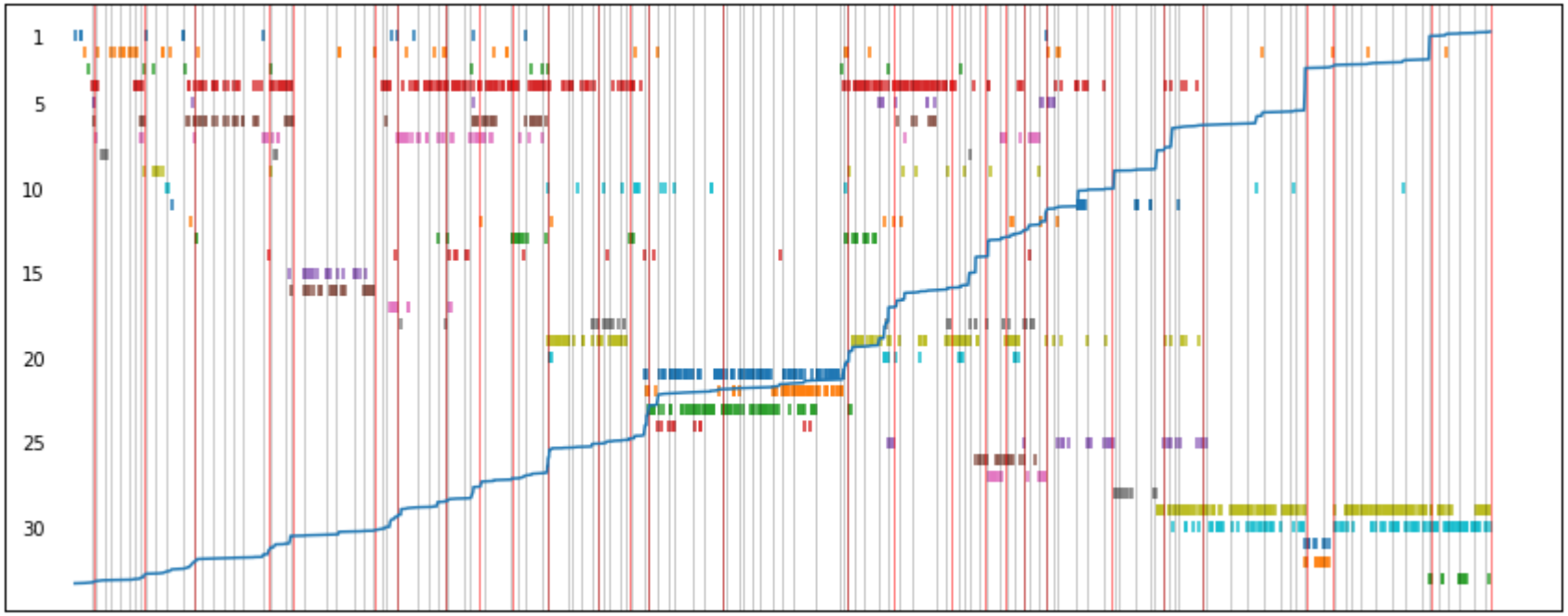


en grand sur le slide suivant

Segmentation du texte



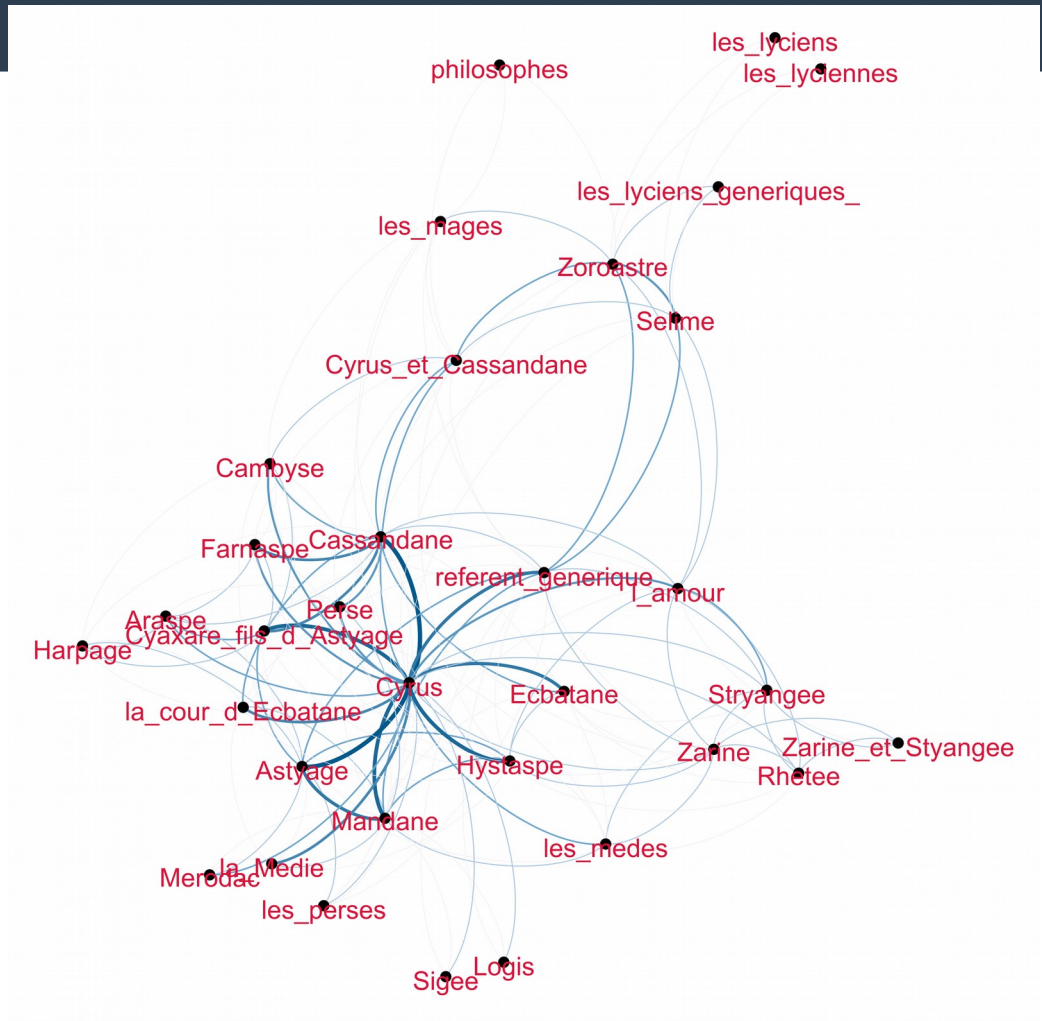
Segments vs. paragraphes



Comment lire les segments?

- **relation entre les segments du texte**
  - association des référents en fonction de leur rencontre dans le texte:
  - la relation est d’autant plus forte que les référents se retrouvent dans les mêmes segments

Comment lire les segments?



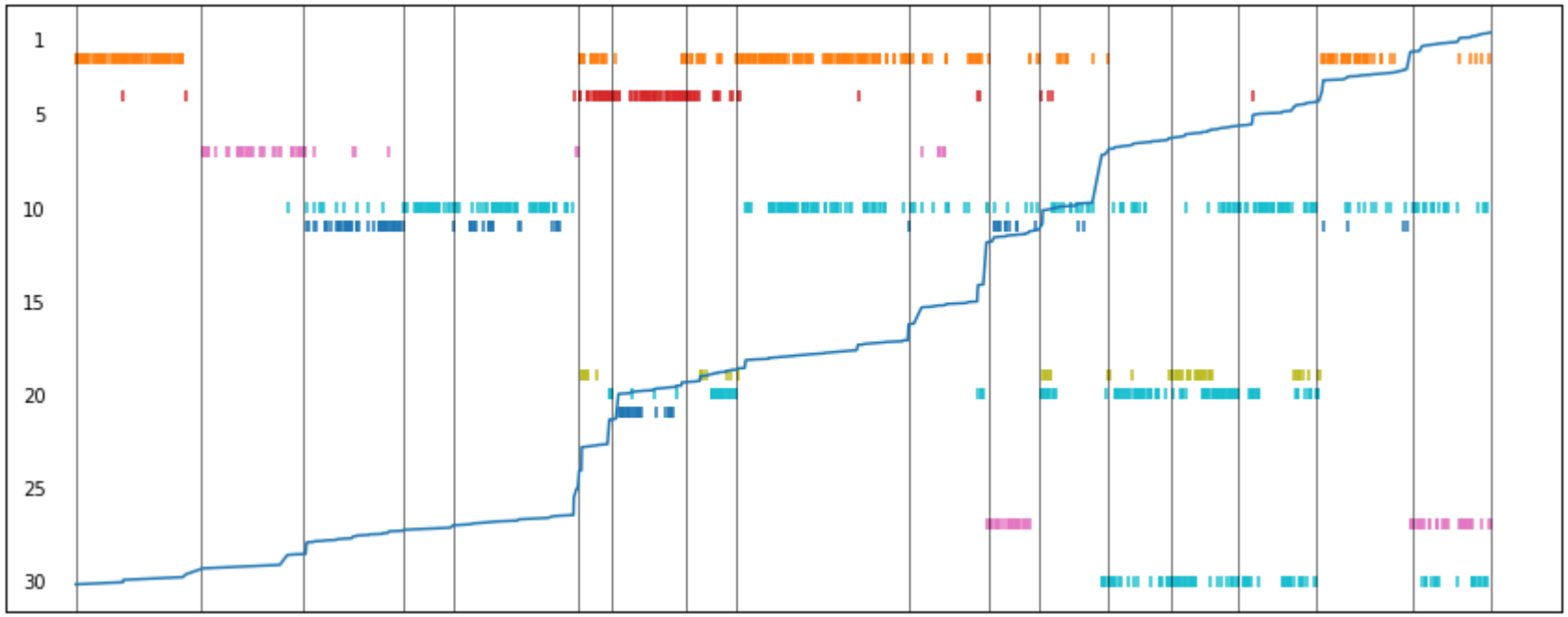
Comment lire les segments?

- **segments établis sur la base de la présence de certains référents:**
  - segments caractérisés par un ou deux référents principaux présents *ensemble*
- **ex.:**
  - épisode de Zoroastre et Sélime
  - épisode de Strangée et Zarine
  - etc.

Perspectives

Perspectives

- **d’autres textes… par ex. Nemoville**



Perspectives

- **autres points à envisager**
  - essai de correspondance avec d’autres façon de segmenter les textes, par exemple avec le lexique
  - étude des transitions
  - toutes les chaînes (et pas seulement celles de 10 maillons ou plus) -> comment déterminer le seuil?

Deux autres indicateurs intéressants:

enchevêtrements
et noms vs pronoms

Autres indicateurs

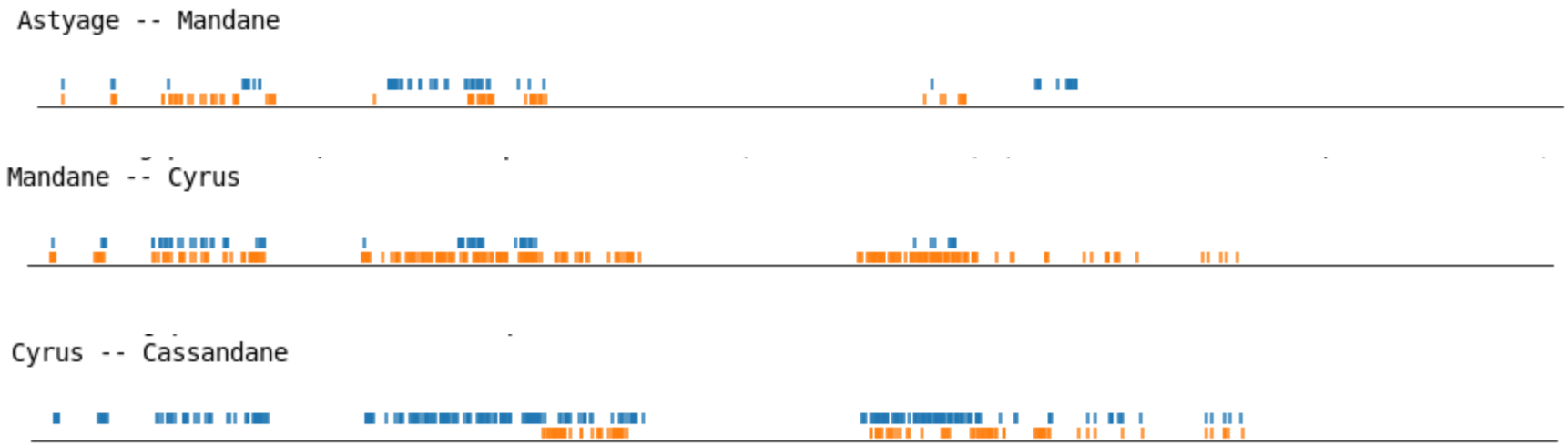
1. les enchevêtrements de Lafon

L’enchevêtrement

- **Lafon, “Dépouillements et Statistiques en lexicométrie”, 1984**
  - indicateur lexicométrique qui a été adapté aux chaînes de référence
- **calcul de l’enchevêtrement *via* deux indicateurs:**
  - quel est le degré d’enchevêtrement?
  - quelle est la distance entre les paires de mentions enchevêtrées

L’enchevêtrement

- **utile pour étudier la cohabitation des chaînes, et donc aussi la structuration textuelle**



L’enchevêtrement

- **Deux chaînes enchevêtrées**

- **Deux chaînes non enchevêtrées**

L’enchevêtrement: exemples

Ecbatane -- les_perses
Prob of fg pairs = 3, out of 21 pairs: 0.995220 (ratio: 0.489552) (distance: z=0.050426, mean=0.022584)
Ecbatane -- les_perses

Ecbatane -- les_medes
Prob of fg pairs = 5, out of 24 pairs: 0.918822 (ratio: 0.723507) (distance: z=26.357051, mean=0.236209)
Ecbatane -- les_medes

Mandane -- Cyrus
Prob of fg pairs = 48, out of 332 pairs: 0.993634 (ratio: 0.495702) (distance: z=41.515202, mean=0.007775)
Mandane -- Cyrus

Perse -- les_medes
Prob of fg pairs = 4, out of 22 pairs: 0.966819 (ratio: 0.628765) (distance: z=11.573982, mean=0.138097)
Perse -- les_medes

Cyrus -- Mandane
Prob of fg pairs = 48, out of 332 pairs: 0.993634 (ratio: 0.495702) (distance: z=12.647588, mean=0.002221)
Cyrus -- Mandane

Rhetee -- Stryangee
Prob of fg pairs = 22, out of 152 pairs: 0.999992 (ratio: 0.496269) (distance: z=-0.337661, mean=0.001481)
Rhetee -- Stryangee

L’enchevêtrement

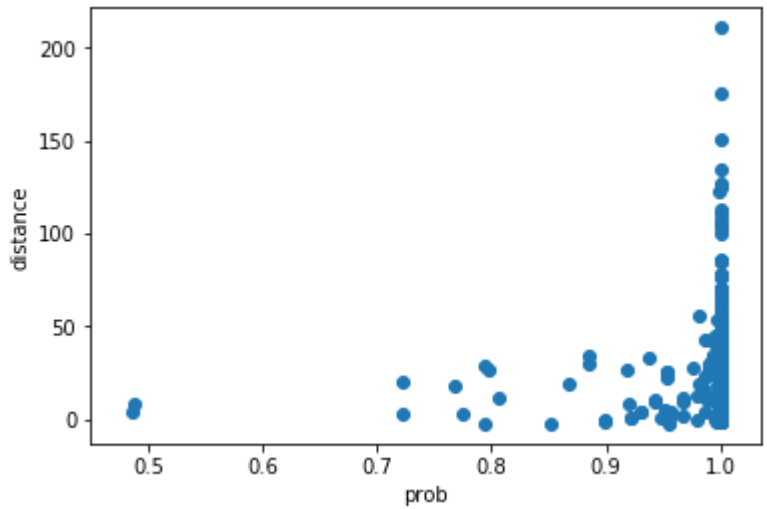
- **difficultés soulevées par Lafon**
  - certaines chaînes ont un haut degré d’enchevêtrement, mais ne sont pas liées car trop éloignées

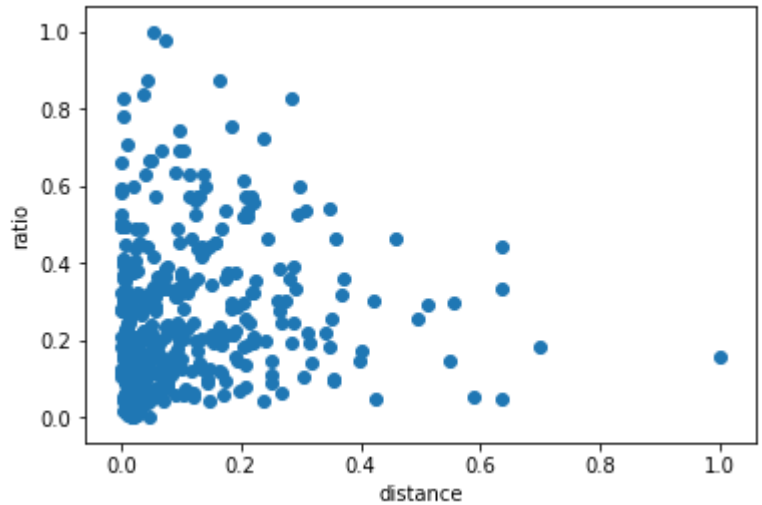
L’enchevêtrement

- **difficultés soulevées par Lafon**
  - d’autres ont un faible degré d’enchevêtrement alors que certaines parties sont très enchevêtrées

L’enchevêtrement

- **difficultés soulevées par Lafon**
  - difficile de concilier les deux chiffres (enchevêtrement et distance) en un indicateur unique





L’enchevêtrement

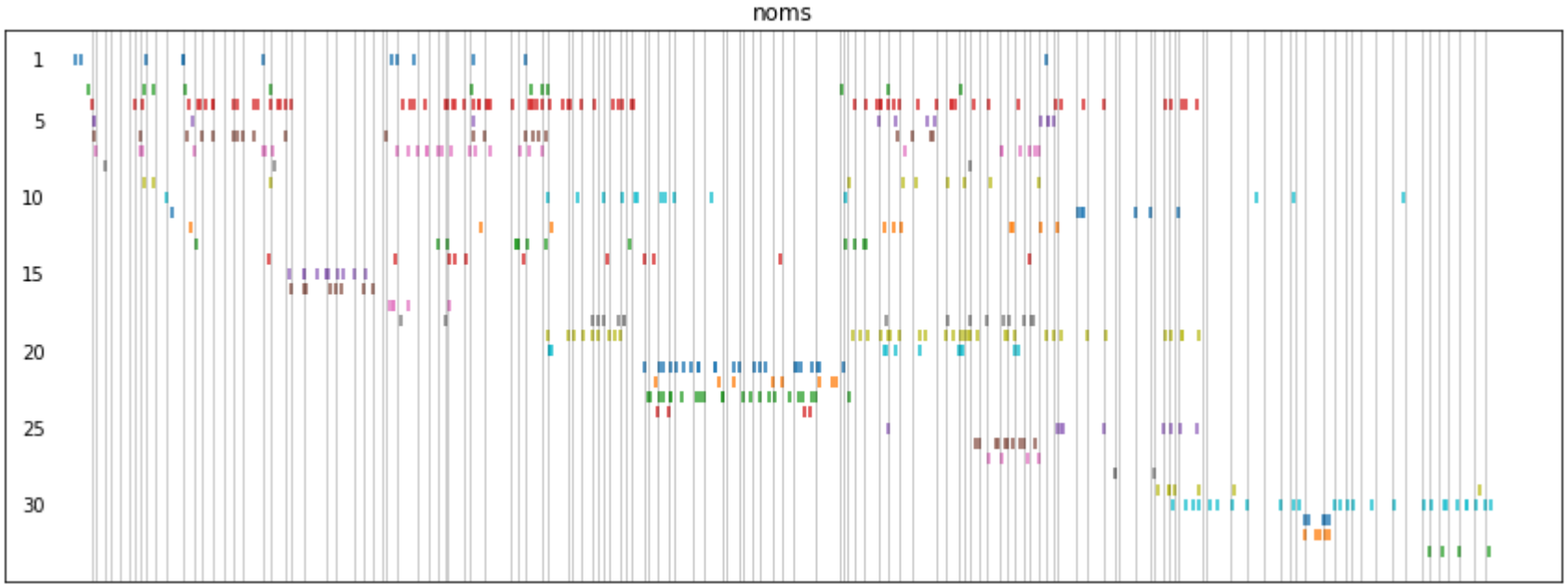
- **utilité**
  - repérage des chaînes successives, non enchevêtrées (non parallèles)

  - n’appliquer le calcul d’enchevêtrement que sur:
    - des chaînes rafaleuses
    - ou des sous-chaînes parallèles

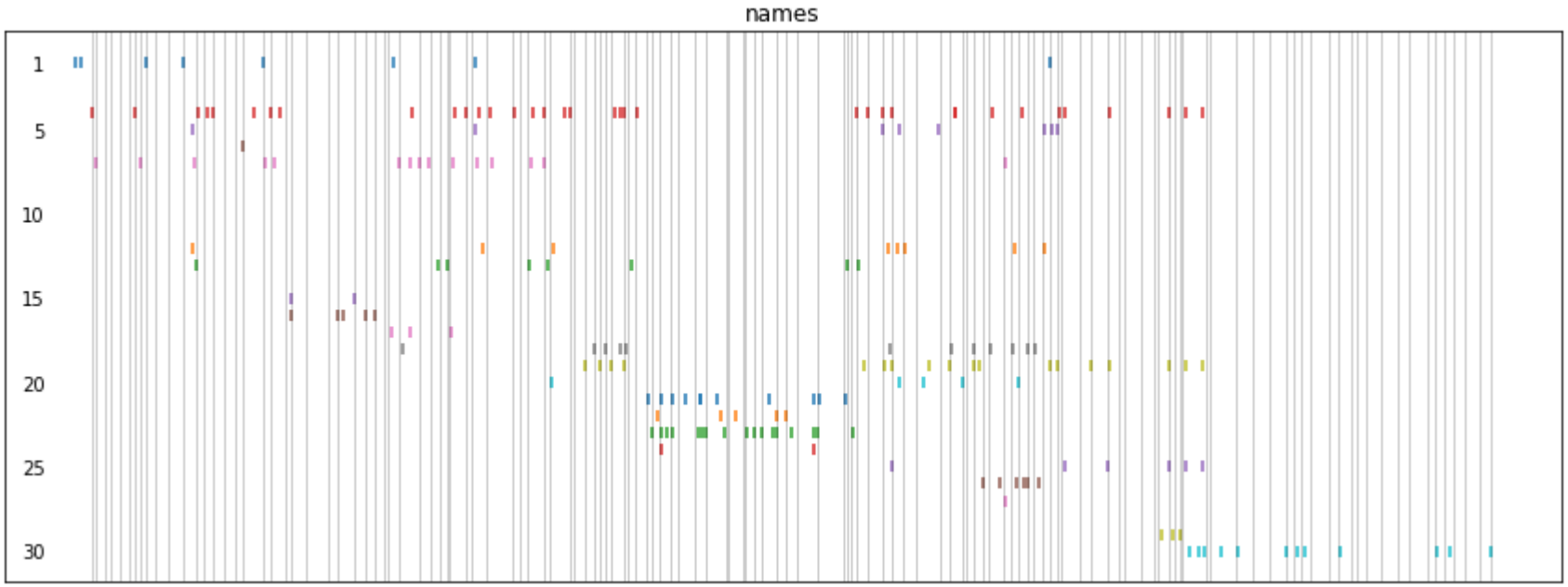
Autres indicateurs

2. noms vs pronoms

Les noms

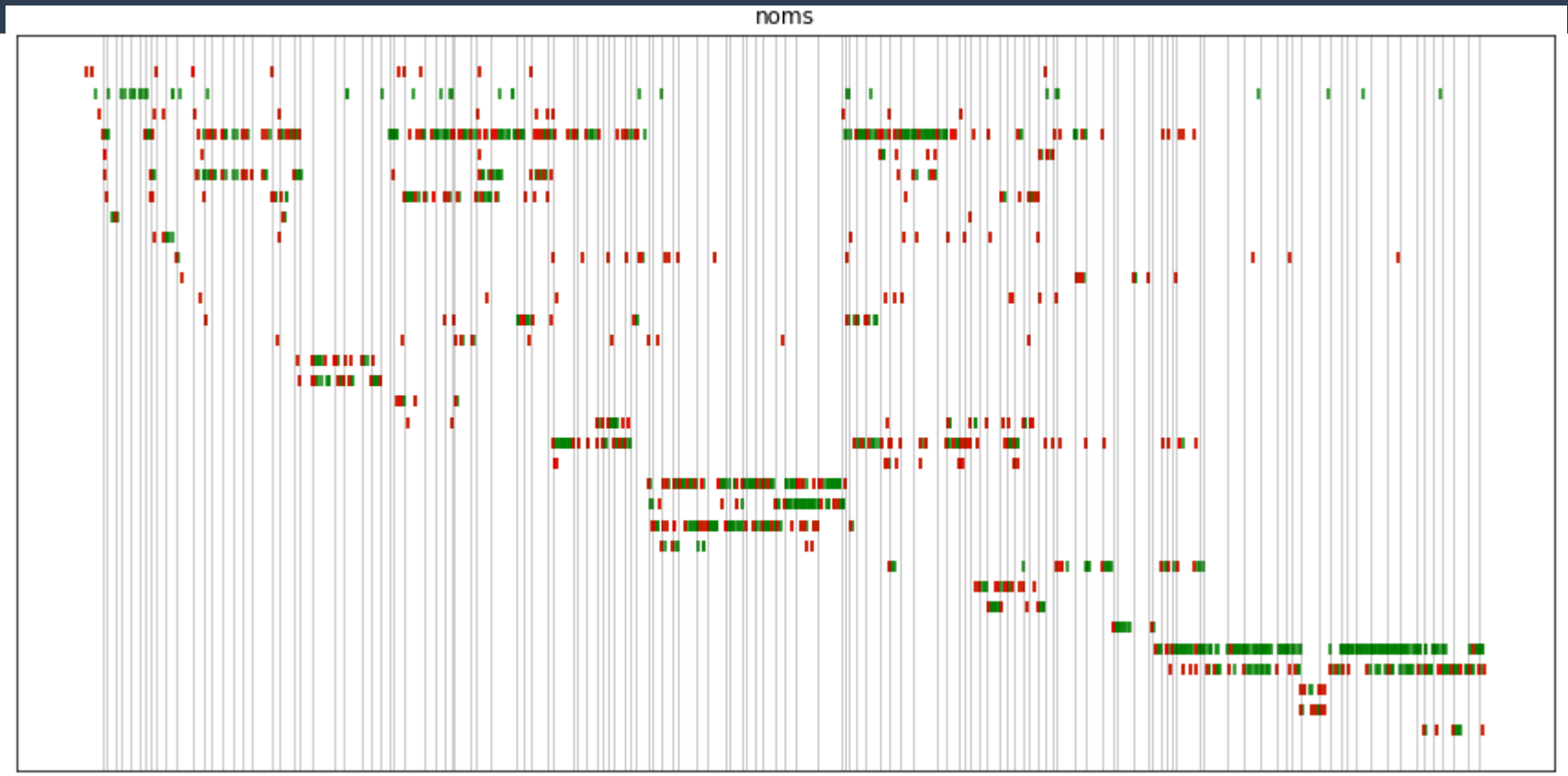


Les noms propres

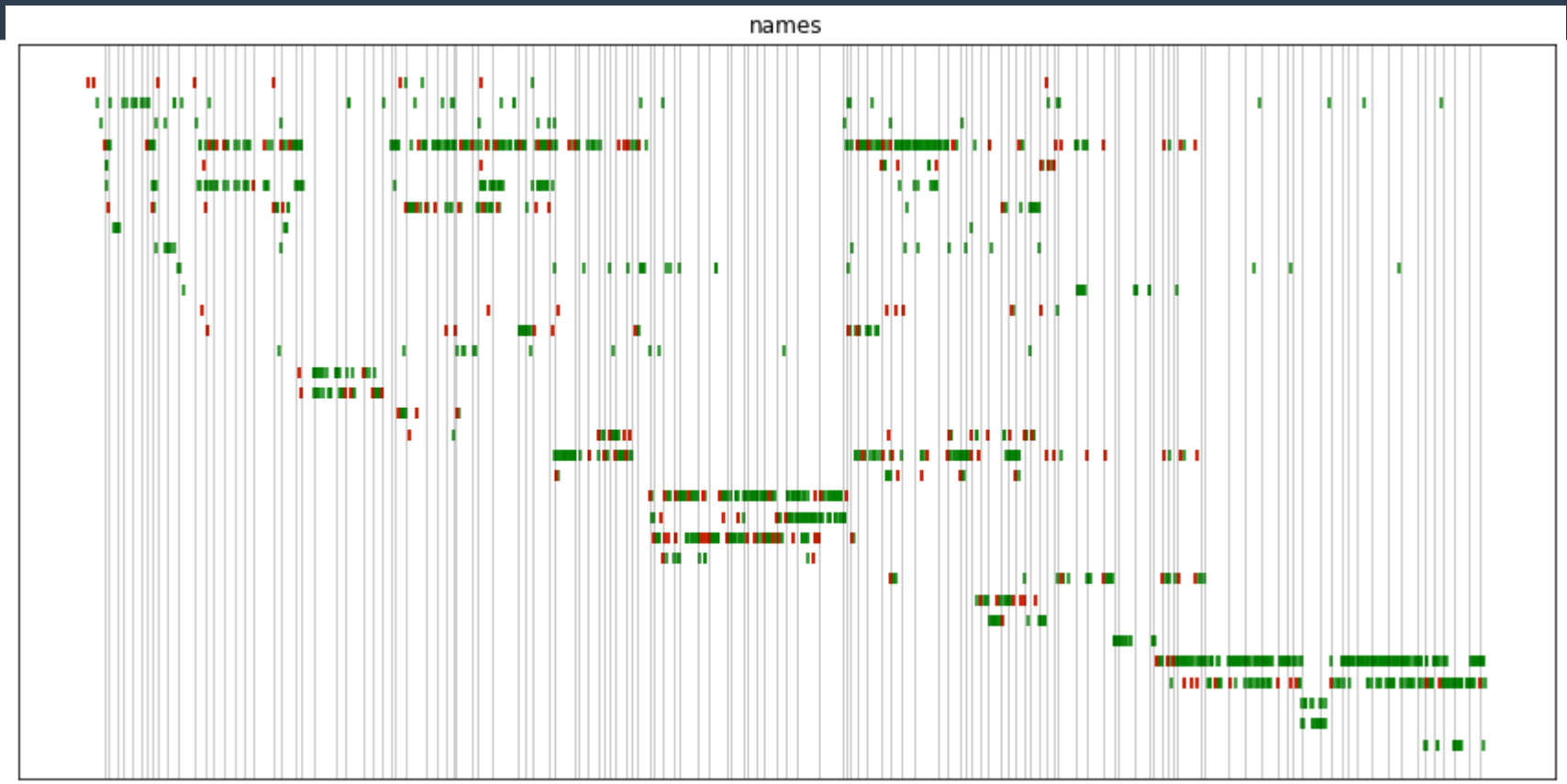


Les pronoms

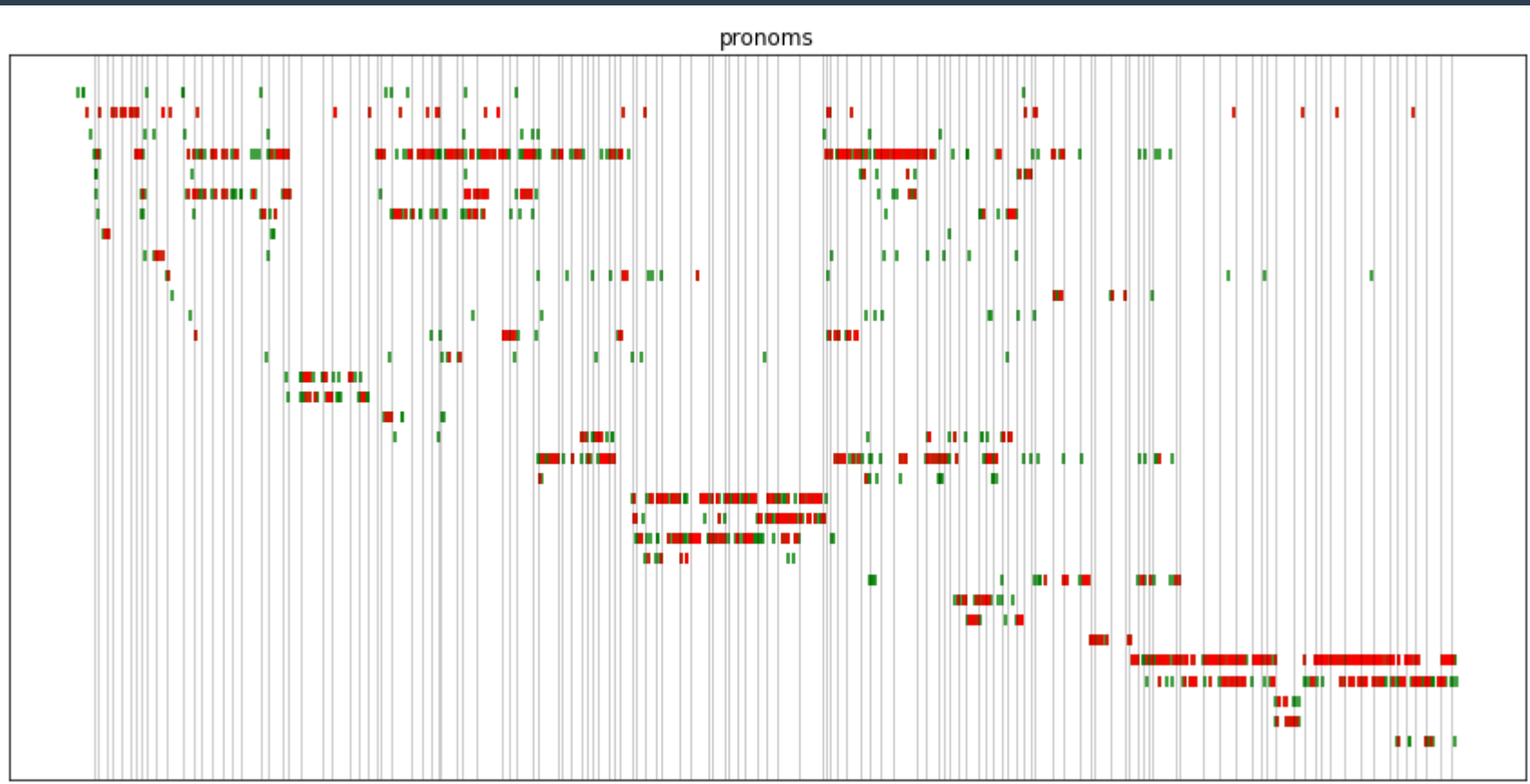
Noms (rouge) vs pronoms et dét. (vert)



Noms propres (rouge) vs le reste (vert)



Pronoms (rouge) vs le reste (vert)



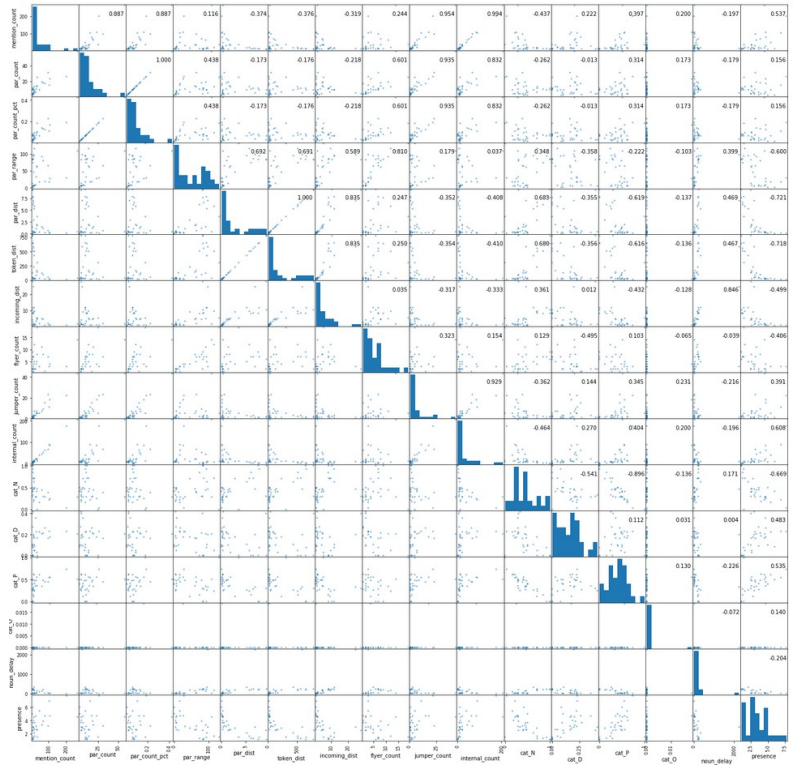
Noms vs. pronoms

- **un domaine à explorer**
- **difficile de tirer des conclusions à partir des graphiques**

Marché aux indicateurs

Difficultés

- facile de trouver des indicateurs
- difficile de repérer ceux qui sont les plus pertinents et les plus efficaces



Rang des mentions dans le paragraphes

- **les mentions de rang 1/2/etc. sont les premières/deuxièmes/etc. mentions de chaque chaîne dans chaque paragraphe**
- **donc: nombre de mentions de rang 1**

  = nb de chaînes * nb de paragraphes
- **ex:**

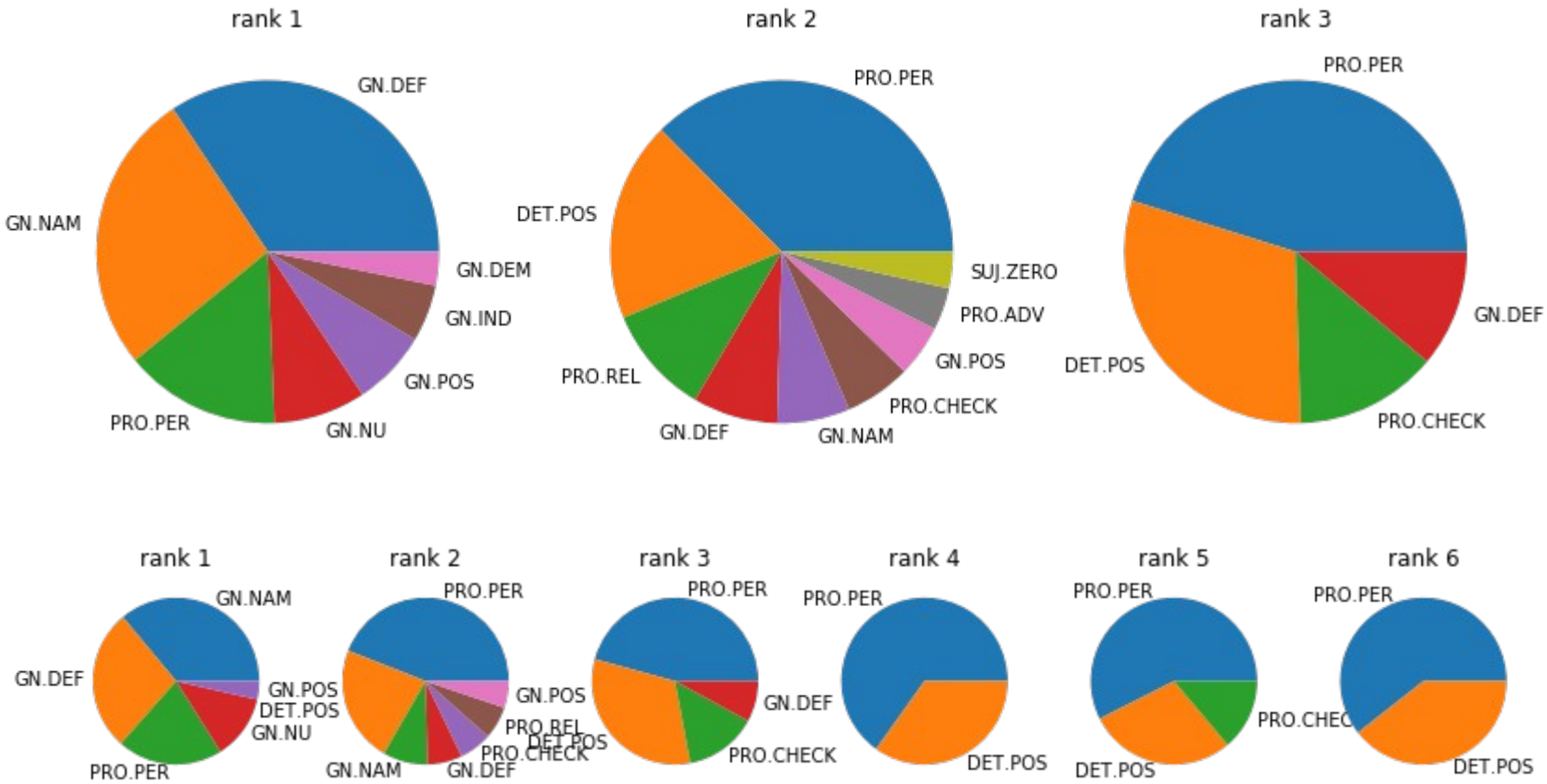
1 2 1 2

par1: le chat... il... la chienne... elle...

par2: Félix... il... la chienne... elle...

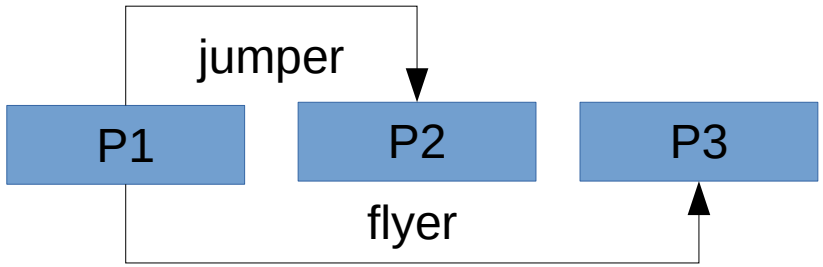
1 2 1 2

Rang des mentions dans le paragraphes



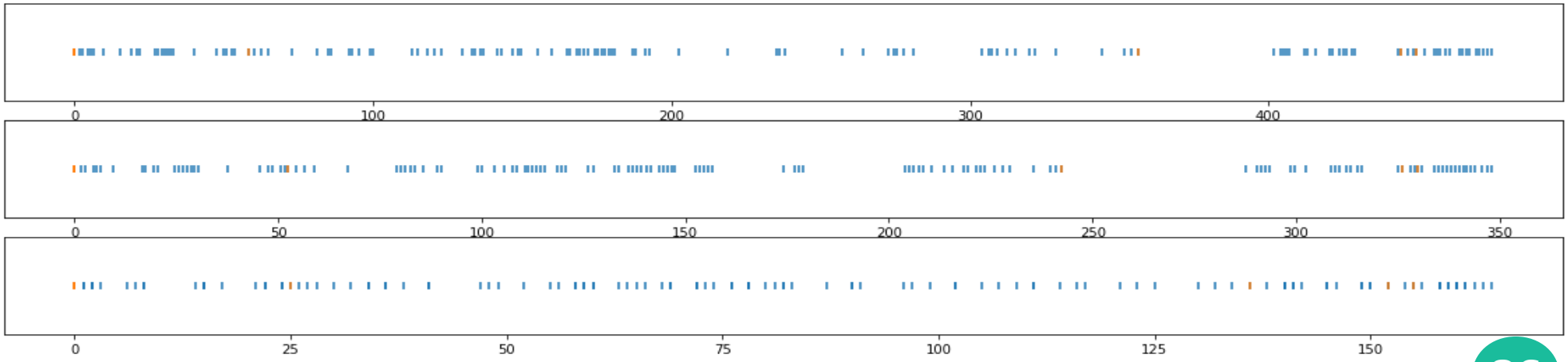
Jumpers et flyers

- **jumper: chaîne qui saute d’un paragraphe au suivant**
- **flyer: chaîne qui survole au moins un paragraphe sans s’y arrêter**



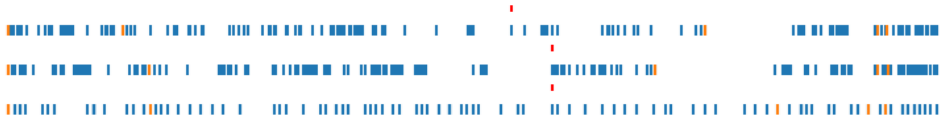
Jumpers et flyers

- **étude de l'effet des flyers sur les paragraphes**
- **représentation:**
  - chaque trait est un paragraphe
  - les espaces blancs représentent la distance moyenne (ou médiane, ou le nombre) des flyers entrant dans le paragraphe



Jumpers and flyers

Paragraph 79



- incoming dist (mean): 19.571428571428573
- incoming dist (median): 25.0
- incoming flyers: 5.0
- incoming jumpers: 2.0
- main chain: Cambyse
- main chain density: 2.25

incoming flyers:

- Cambyse (39)
- Farnaspe (33)
- Perse (33)
- Cyaxare_fils_d_Astyage (25)
- Ecbatane (5)

incoming jumpers:

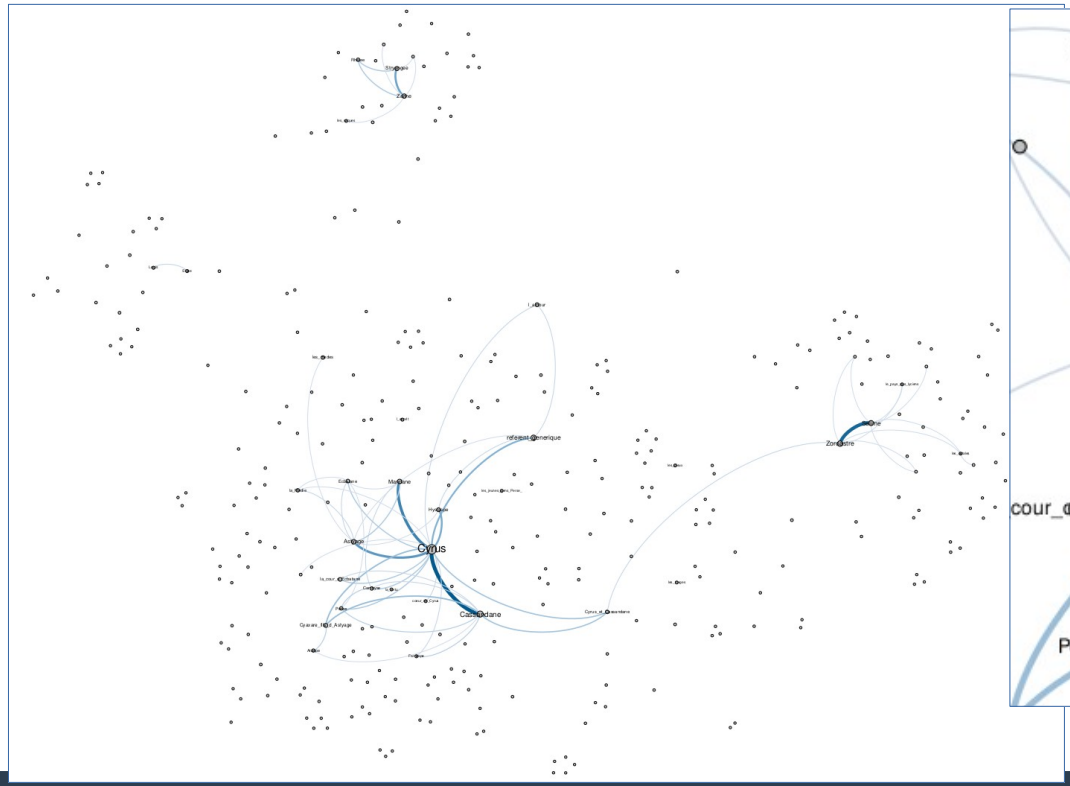
- Cassandane
- Cyrus

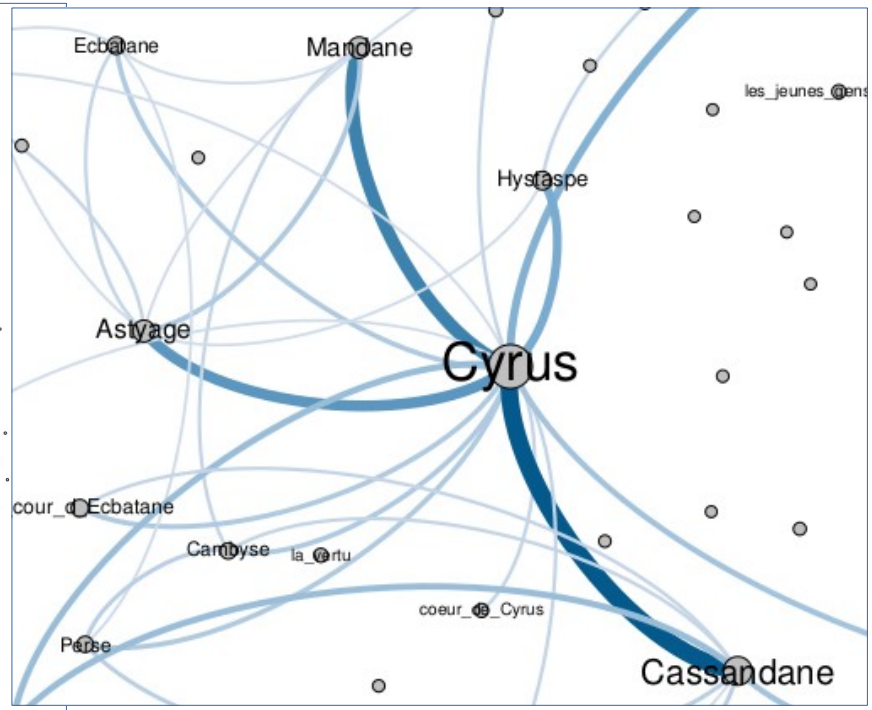
new chains:

Cependant Cambyse apprit l' amour de Cyrus pour Cassandane ; mais ayant d' autres vûes pour son fils , qui s' accordoient mieux avec sa politique , il le rappelle en Perse . Farnaspe qui étoit toujours à la cour de Cambyse , fut instruit en même temps des sentimens de Cyaxare . Le satrape ambitieux flatté par cette alliance , ordonna à sa fille de rester à Ecbatane .

Réseaux de référents

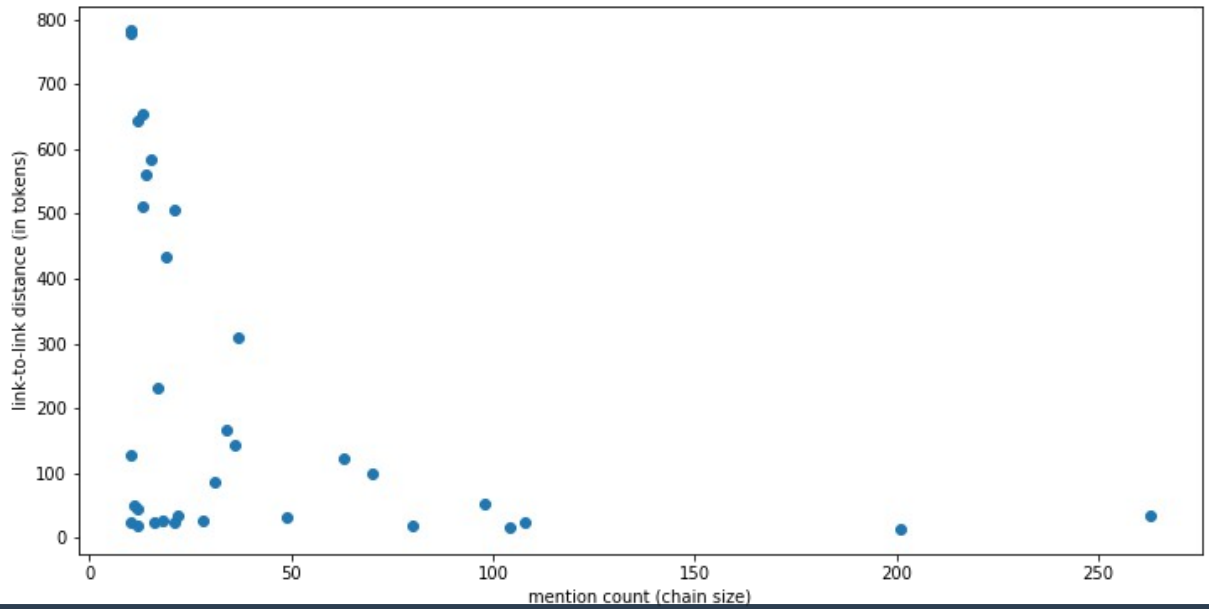
- **nombre de paragraphes dans lesquels les référents se retrouvent**



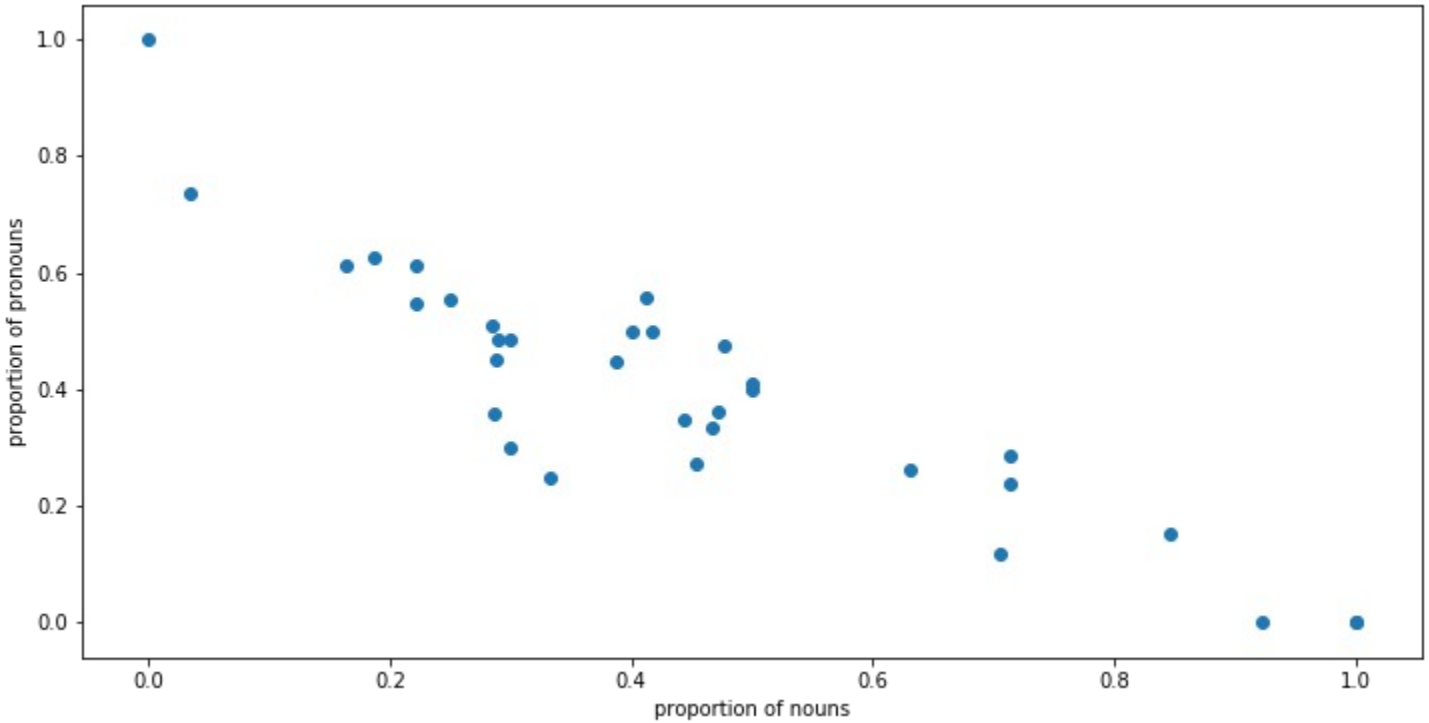


Taille vs distance intermaillonnaire

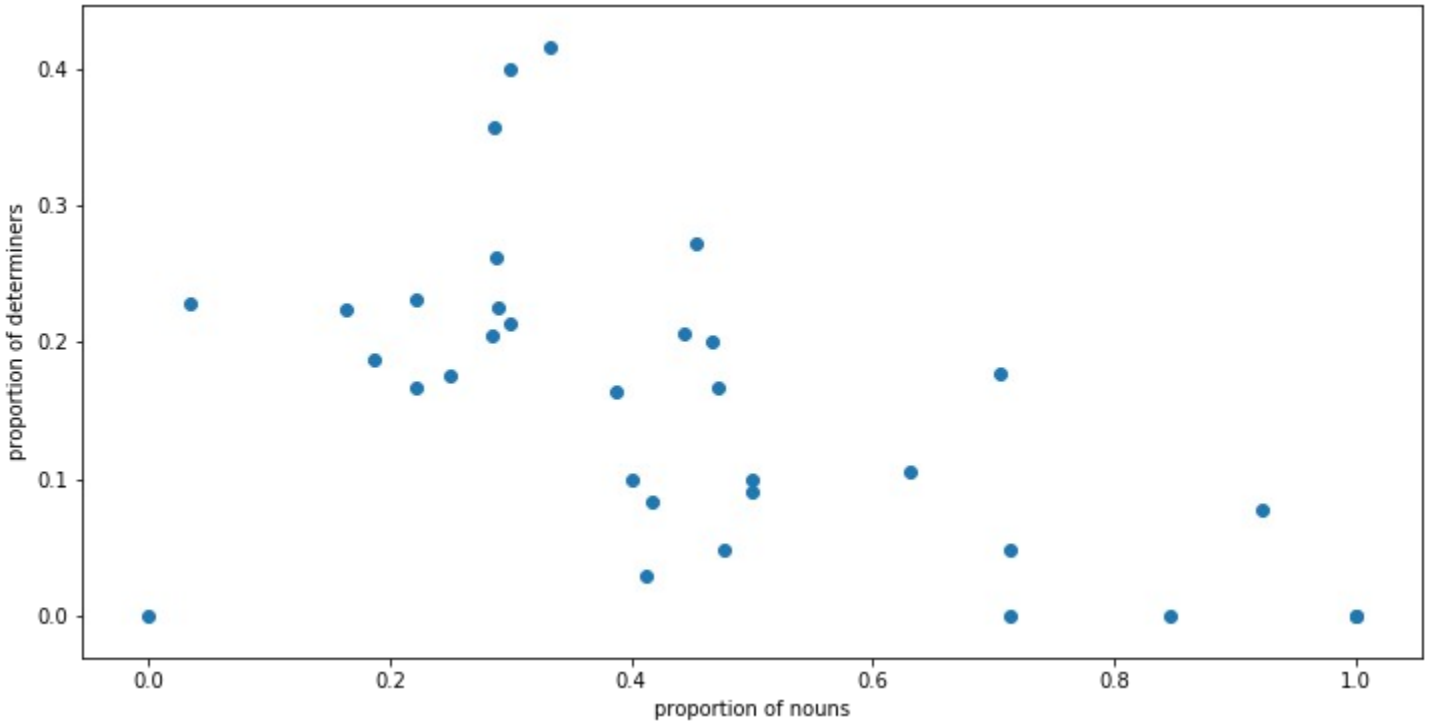
- **deux types de chaînes:**
  - avec beaucoup de maillons et resserrées (cf rafales)
  - avec peu de maillons mais distendues (cf régulières)



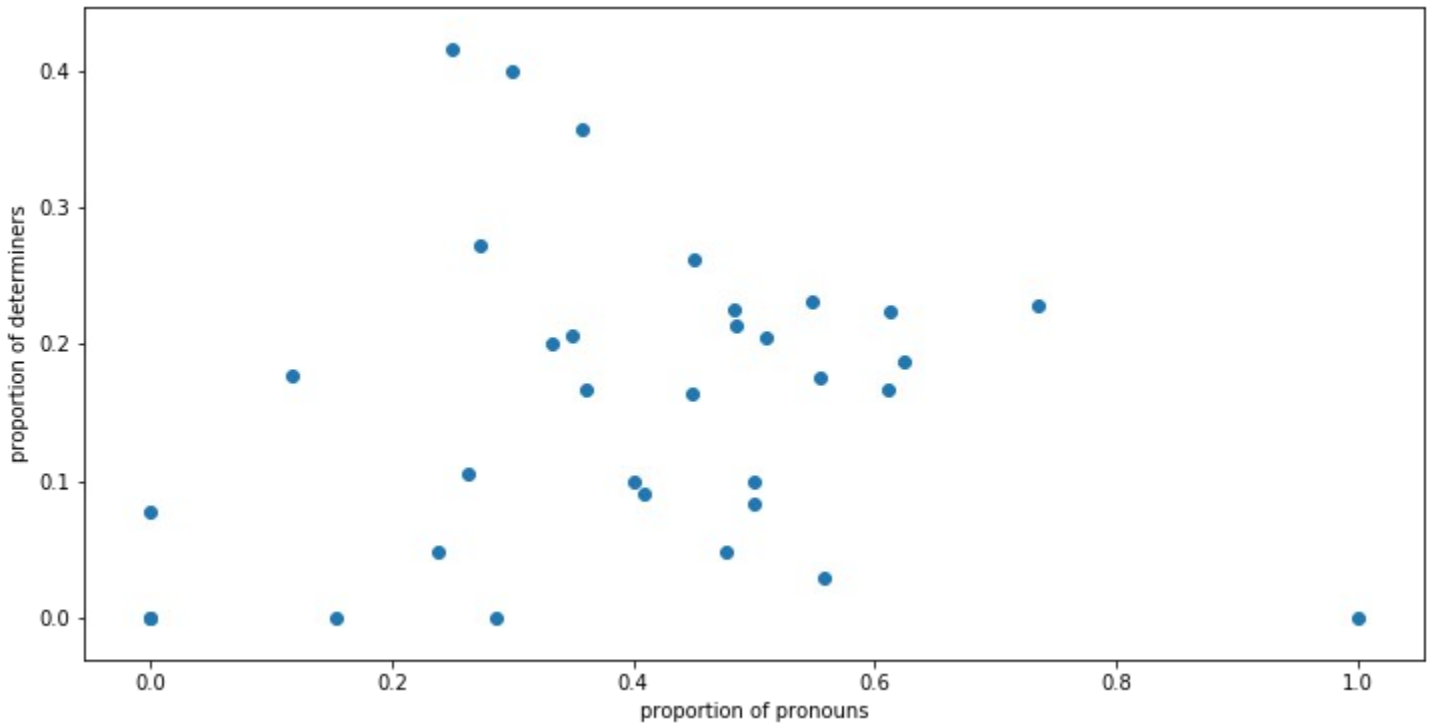
Proportions de noms vs. Prop. de pronoms



Proportions de noms vs. Prop. de det.



Proportions de pronoms vs. Prop. de det.



Merci de votre attention!