Exploiter les données

Introduction à la textométrie avec TXM

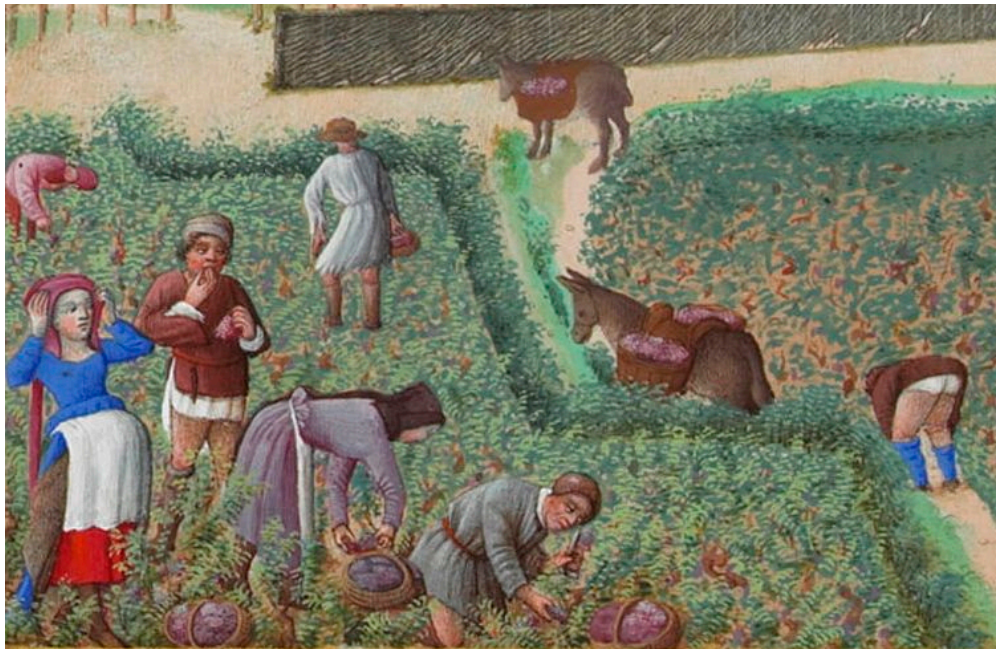
Jean-Baptiste Camps & Simon Gabay

Univ. de Neuchâtel

Formation en philologie numérique :
encoder, exploiter, diffuser
12-16 février 2018

De la production des données à l'exploitation

Ceux-ci collectent les données



Ceux-là les exploitent

Objectifs

- S'initier à la textométrie,
  - créer des corpus ;
  - les interroger ;
  - faire (un peu) d'analyse quantitative.
- dans le cadre d'un logiciel "tout en un" et convivial, TXM [Heiden et al., 2010],
- sur des cas tirés de la littérature du XVII^{e} siècle.

La textométrie selon [Pincemin et al., 2008]

La textométrie développe les possibilités de consultation et d'analyse de corpus textuels en faisant appel à des décomptes et des modélisations statistiques et en combinant aux possibilités de repérage d'occurrences des calculs de tri, de sélection et de réorganisation statistique.

TXM : un logiciel de textométrie

`http://textometrie.ens-lyon.fr/`
Base de français médiéval :
`http://txm.bfm-corpus.org/`.

- Logiciel libre et multiplateforme ;
- développé à l'ÉNS-LSH de Lyon ;
- dévoué à la textométrie ;
- repose sur des technologies de référence :
  - XML/TEI pour les données ;
  - R pour l'analyse statistique ;
  - CQP pour l'interrogation de corpus ;
  - `TreeTagger` pour l'annotation.
- existe en version
  - bureau ;
  - serveur.

Installer TreeTagger dans TXM

Il existe plusieurs outils d'annotation linguistique (lemmatisation, étiquetage morpho-syntaxique, syntaxique…) qui peuvent s'utiliser seuls (`TreeTagger`, `Wapiti`, `Pandora`, Marmot, `Mate Tools`, …) pour préparer un corpus.
`TXM` permet en outre d'utiliser `TreeTagger` pour annoter les corpora lors de leur création.
Aller dans aide / installer TreeTagger et suivre les étapes présentées pour votre système d'exploitation.

Plan

Plan

Création d'un corpus à partir de notre édition d'*Andromaque*

1. Fichier, importer, import XML/W + CSV ;
2. sélectionner le dossier avec les sources et remplir les paramètres du corpus ;
3. lancer la création du corpus.

Premières fonctionnalités

1. Consulter la description du corpus ;
2. parcourir l'édition ;
3. regarder le lexique ;
4. ouvrir l'index, y chercher les occurrences de 'Seigneur' ;
    1. clic-droit, envoyer vers les concordances ;
        1. double-clic sur une occurrence pour aller au texte ;
    2. clic-droit, envoyer vers les cooccurrents ;
        1. aller d'un cooccurrent aux concordances, puis au texte
    3. clic-droit, envoyer vers la progression ;

Partitions et quelques éléments descriptifs

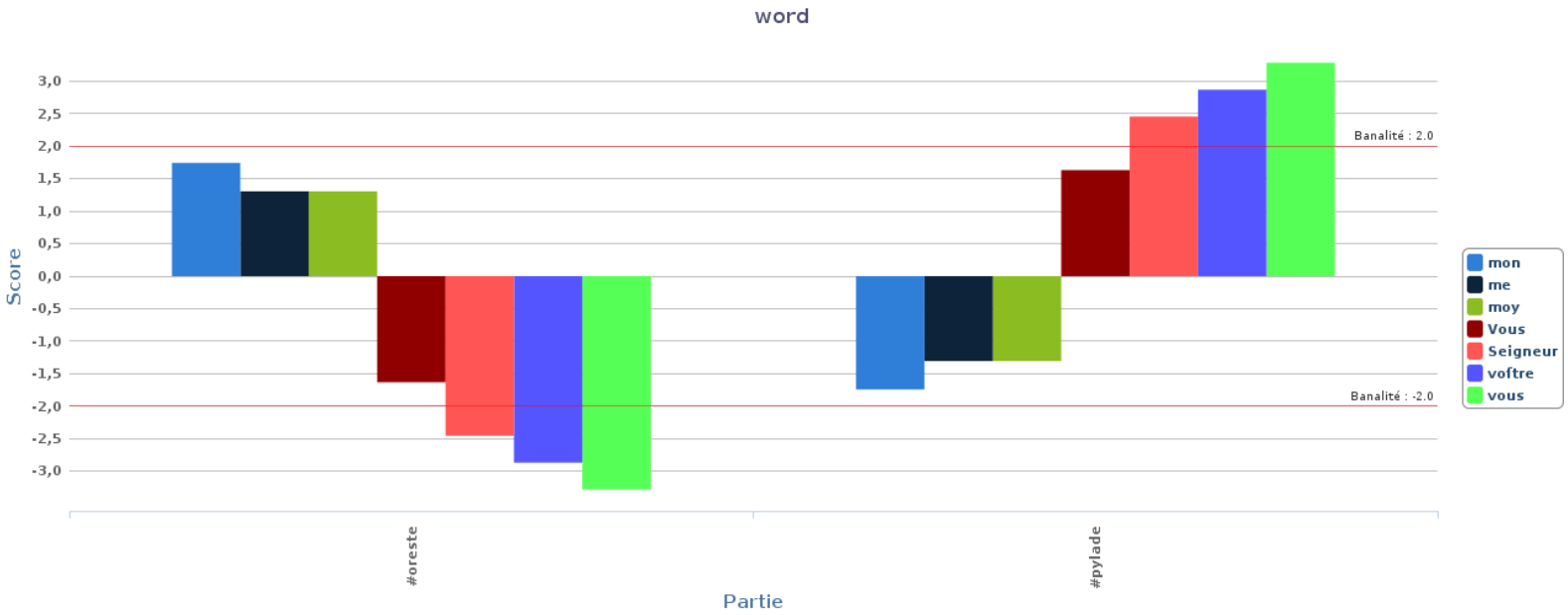
1. créer une partition, en sélectionnant la structure `sp` et l'attribut `who` ;
2. consulter les dimensions ;
3. créer une table lexicale, expérimenter avec les tris, la fusion ou suppression des colonnes, etc.

Statistiques de base

À partir de la table lexicale créée,

1. calculer les spécificités ;
2. quels sont les mots les plus spécifiques d'Oreste ? de Pylade ?
3. en sélectionner quelques uns qui sont pertinents ;
4. calculer le diagramme en bâton des lignes sélectionnées.

Qu'en déduire ?



Prêts à passer aux choses sérieuses ?

Plan

Différents modes d'import

flemmards import (avec ou sans métadonnées complémentaires) depuis

- presse-papier ;
- des fichiers `txt` ;
- traitement de texte.

XML XML/ TEI ou autre ;

spécifiques formats de logiciels de textométrie.

Corpus du jour : un peu de théâtre du XVII^{e} siècle

Corpus constitué pour le cours d'aujourd'hui :

- source : Paul Fièvre, `http://www.theatre-classique.fr/` ;
- documents encodés en XML/TEI (ou dans plusieurs XML/TEI) ;
- corpus de 36 pièces de théâtre en vers du XVII^{e} siècle,
- appartenant à trois genres principaux :
  - comédie,
  - tragédie,
  - tragi-comédie.
- sélectionnées un peu au hasard,
  - mais en essayant de conserver un équilibre entre les genres principaux (comédie, tragédie, tragi-comédie),
  - d'avoir des pièces de longueur similaire (entre 1250 et 2000 vers),
  - un équilibre entre les auteurs (4 pièces par auteur),
  - et entre les générations (4 auteurs par génération).

9 auteurs, 2 générations

G1, c. 1630-1650

- Pierre Du Ryer (fl. 1628-1655) ;
- Georges de Scudéry (fl. 1631-1643).
- Jean de Rotrou (fl. 1635-1649) ;
- Paul Scarron (fl. 1648-1660) ;

G2, c. 1650-1690

- Claude Boyer (fl. 1646-1697) ;
- Thomas Corneille (fl. 1651-1696)
- Molière (fl. 1655-1673)
- Jean Racine (fl. 1664-1691) ;

Et un monstre sacré, **Pierre Corneille** (fl. 1629-1675).

Sources et pré-traitements

- Graphies modernisées (dommage… mais va faire notre affaire dans ce cas précis) ;
- Faut-il supprimer la distinction majuscule/minuscule ?
  - Pour : suppression de biais éditoriaux.
  - Contre : les majuscules peuvent conserver de l'information syntaxique.
- Veut-on garder tout ce qui est extérieur aux répliques (liste des personnages, page de titre, etc.) ?
  - Peut être retiré grâce à la structuration XML/TEI.
- pas de lemmatisation : possibilité de lemmatiser et annoter automatiquement.

Transformations avant l'import

Dossier `xsl`, feuille `source_to_txt.xsl` (2.0), deux sorties :

1. fichier `metadata.csv` : métadonnées des documents extraites automatiquement des fichiers TEI (`teiHeader` et page de titre, `docDate`) ;
2. dossier `txt` : transformation en `txt` des pièces :
   - passage en bas de casse ;
   - suppression du `teiHeader` ;
   - suppression du `castList` et des mentions de personnage, `speaker` ;
   - suppression du `front`, des `docTitle`, `docDate`, `docAuthor`, `docImprint`, `printer`, `performance`, `div[\@type='dedicace']` ;
   - suppression des titres, notes.

Import `txt` + `csv`

1. sélectionner le répertoire de sources `txm_import1_txt` (contenant fichiers `txt` et métadonnées `csv`, corpus tronqué pour gagner un peu de temps) ;
2. paramétrer l'import (nommer le corpus THEATRENEUCHTXT) ;
3. demander la lemmatisation ;
4. vérifier que les métadonnées sont bien comprises ;
5. lancer l'import ;
6. (regarder le log de l'import) ;
7. une fois l'import réussi, jeter un œil à la description, à l'édition, etc.

Import xml/w + csv amélioré

1. sélectionner le répertoire de sources `txm_import2_xml` (contenant fichiers `xml` et métadonnées `csv`, complet) ;
2. paramétrer l'import (nommer le corpus THEATRENEUCHXML) ;
3. demander la lemmatisation ;
4. vérifier que les métadonnées sont bien comprises ;
5. associer l'`xsl` 2.0 de pré-traitement, `import_xml_filtre.xsl` ;
6. lancer l'import ;
7. (regarder le log de l'import) ;
8. une fois l'import réussi, jeter un œil à la description, à l'édition, etc.

Plan

CQP et CQL

Corpus Query Processor

Composant logiciel qui traite des requêtes :
moteur de recherche qui permet de trouver toutes les occurrences correspondant à une requête.

- logiciel libre ;
- développé initialement à l'Univ. de Stuttgart ;
- `http://cwb.sourceforge.net/`.

Corpus Query Language

Langage d'expression de requêtes (cf. SQL, XQuery…).
Une expression CQL est une chaîne de caractères exprimant un motif linguistique (un mot, ou une suite de mots) à partir des valeurs de leurs propriétés (comme la catégorie grammaticale, le lemme, la forme graphique). (voir Manuel de TXM).

Interroger avec TXM : niveau 0

Entrer un mot dans le champ de l'interface `Index`.
Ex. 'seigneur'.

Interroger avec TXM : niveau 1, assistant

Cliquer sur la baguette magique, pour accéder à l'assistant de création de requêtes
Chercher :

- la forme 'seigneur',
- suivie d'un pronom (cf. la doc, `JeuEtiquettesModeleFrancaisTreeTagger.pdf`),
  - astuce : commence par 'PRO'
- suivi de n'importe quelle forme du lemme 'être'.

Interroger avec TXM : niveau 2, un peu de CQL

En CQL, la requête précédente correspond à :

```
[word="seigneur"][frpos="PRO.*"] [frlemma="être"]
```

On se lance :

1. Modifier la requête pour permettre un mot, quel qu'il soit, entre seigneur et le pronom ;
2. la modifier, pour limiter aux cas où ce mot est une virgule ;
3. l'éditer pour étendre aux cas où ce mot est une virgule OU un point d'interrogation.

Interroger avec TXM : niveau 3, CQL (plus) avancé

Ignorer :

%c casse, ex.
`[word="état"%c]`

%d diacritiques, ex.
`[word="état"%d]`

%d les deux, ex.,
`[word="état"%cd]`

Opérateurs :

= égal

!= différent

| ou

& et

() priorité des opérations

Quantificateurs

mot mot une seule fois (1) ;

mot+ mot une seule fois ou plus (1…n) ;

mot ? mot 0 ou une fois (0…1) ;

mot* mot 0 fois ou plus (0…n) ;

mot{2,4} mot entre 2 et 4 fois (2…4) ;

Interroger avec TXM : niveau 3, (suite)

Échapper des caractères spéciaux

Pour entrer '?', '*', '.', '+', '|', '&', ..., qui sont des caractères spéciaux,
les **faire précéder d'une barre oblique inverse**. Ex., \?

Classes de caractères

. n'importe quel caractère ;

[mn] un m ou un n ;

[a-z] une minuscule non accentuée ;

[^a-z] tout sauf ... ;

\d un chiffre ;

\s un caractère d'espacement ;

\w un caractère de mot ;

\D tout sauf un chiffre ;

\S tout sauf un car. d'espacement ;

\W tout sauf un car. de mot ;

Interroger avec TXM : niveau 3, exercices

1. un mot contenant un chiffre.
2. une forme commençant par 'a' et finissant par 'er' ;
3. une forme de deux lettres : une voyelle et 'h' ;
4. 'seigneur' ou 'dame', suivi ou non d'un mot et suivi d'un pronom (n'importe quel type de pronom).
5. la forme 'je', suivie d'entre 2 et 4 mots, et d'une virgule.

Interroger avec TXM : niveau 4, CQL *hardcore*

Expressions régulières plus avancées

Classes de caractères Unicode

`\p{Lu}` une majuscule (au sens de la propriété Unicode) ;

`\p{P}` une majuscule (au sens de la propriété Unicode) ;

etc.

N.B. : Toutes les PCRE (Perl-Compatible Regular Expressions) sont disponibles dans CQL. Voir la doc.

Instructions de CQL

On peut limiter la zone de recherche en utilisant `within`. Ex.

```
[word="et"][]*[word="je"] within 10
```

'et … je', dans une limite de 10 mots

```
[word="et"][]*[word="je"] within l
```

'et … je', dans un vers.

Interroger avec TXM : niveau 4, suite

Utiliser les propriétés de structure

Possible d'utiliser les propriétés de structure et XML pour préciser les requêtes.

```
<l> [frpos="VER.*"] [] * </l>
```

Vers commençant par un verbe.

```
<l>  [] * [word=".*uite"] </l>
```

Vers ayant 'uite' à la rime.

Want more ? La doc de CQL est pour vous, ainsi que le manuel de TXM.

Interroger avec TXM : niveau 4, exercices

1. une forme contenant de la ponctuation ;
2. un vers d'entre 4 et 5 mots ;
3. un vers débutant par un pronom personnel (`PRO:PER`) débutant par 't' ;
4. un vers se terminant par '-ron' (suivi ou non d'une consonne).

Plan

Deux approches : la forme (style) ou le fond (sémantique) ?

stylométrie attribution, datation, localisation des textes.

- information graphique, flexionnelle, etc.
- mots les plus fréquents (mots-outils, mots-vides), moins sensibles aux variations intentionnelles de leurs auteurs (genre, sujet, etc.).

approche sémantique (lexicométrie, lecture distante,...), à peu près l'inverse de la précédente :

- lemmes ;
- cooccurrents ;
- mots plus rares.

Un peu d'exploration du corpus

Vérifions tout d'abord si le corpus est bien équilibré, et étudions les distributions par :

- genre ;
- auteur ;
- période
- taille des textes.

Pour ce faire :

1. créer une partition pour chacune de ces questions, grâce aux métadonnées de `text` ;
2. observer les `dimensions`.

Un peu d'exploration des genres

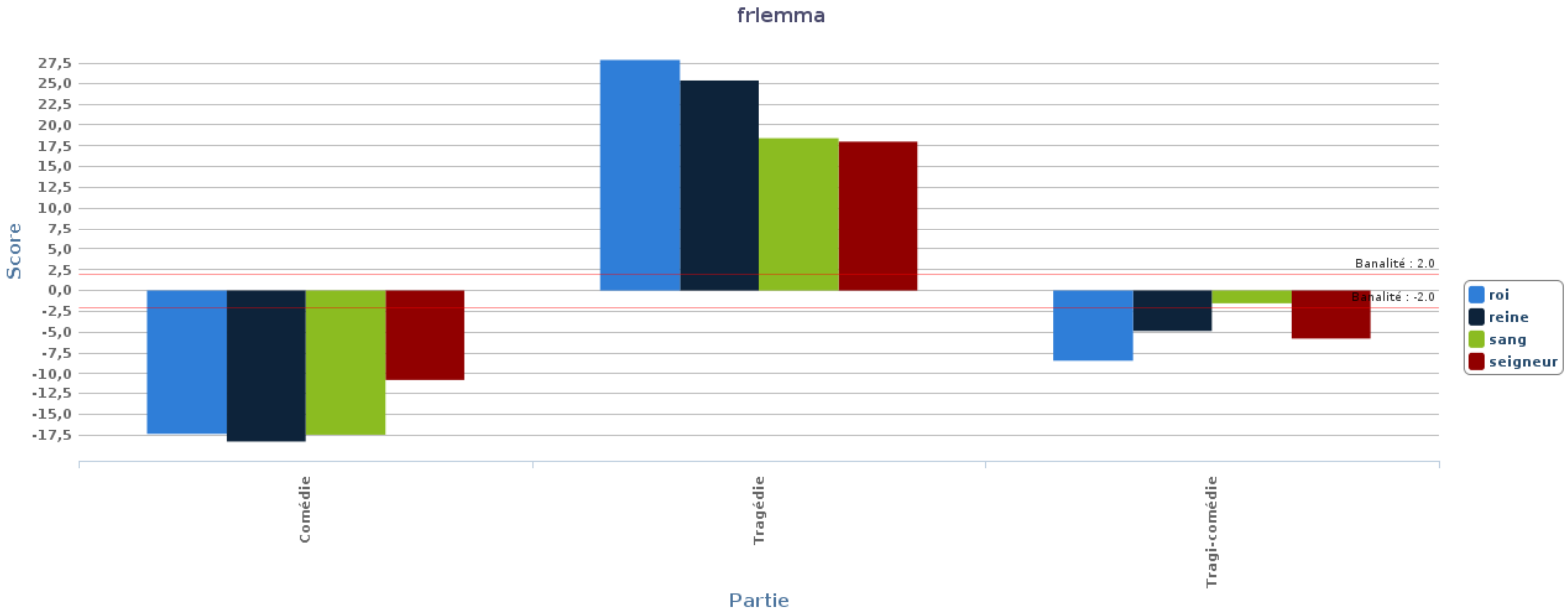
Thématique

1. créer une table lexicale (idéalement, des lemmes) de la partition genre, des 500 formes les plus fréquents ;
2. la traiter pour retirer les mots vides ;
3. calculer les spécificités, sélectionner pour mettre en valeur les particularités de la tragédie, créer le graphique ;
4. calculer la classification ;
5. calculer l'AFC.

Stylistique

1. créer une table lexicale des formes de la partition genre, des 200 formes les plus fréquentes ;
2. la traiter pour retirer les mots porteurs de sens ;
3. calculer la classification ;
4. calculer l'AFC.

(Quelques) lemmes spécifiques de la tragédie



Problématiques autoriales

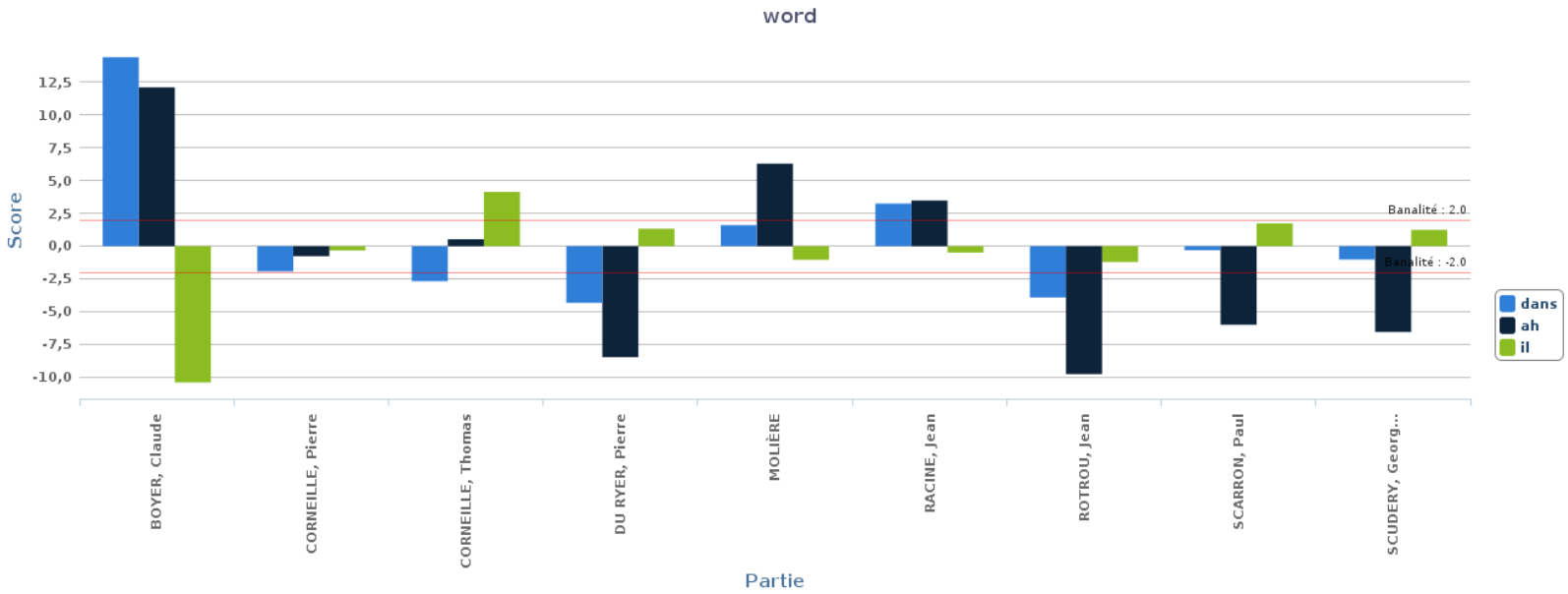
Passons maintenant aux choses sérieuses et aux questions d'attribution.

Étape 1 : observer la répartition des auteurs

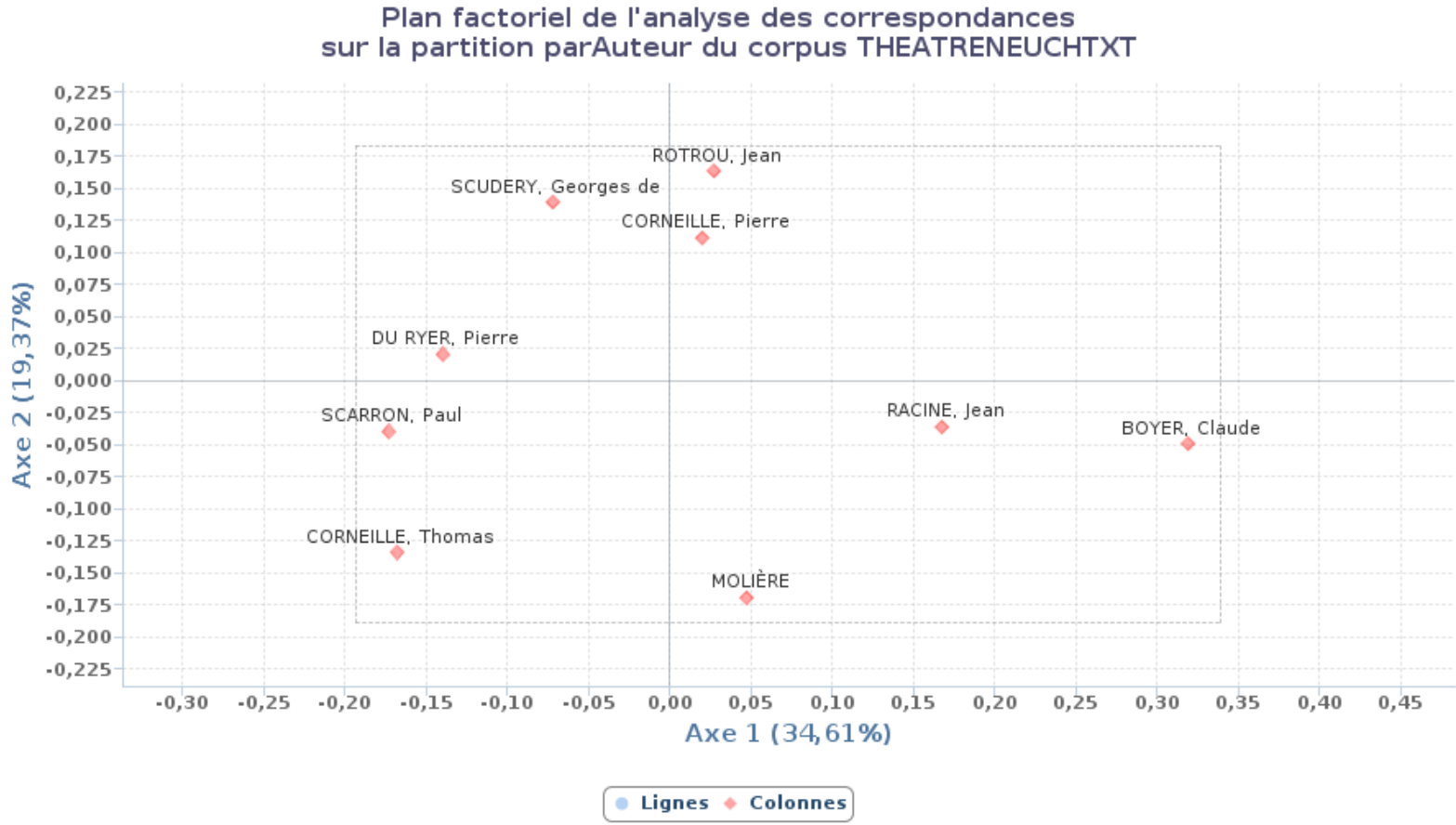
Sur la partition par auteurs,

1. Créer une table lexicale des 200 formes les plus fréquentes, retirer les mots porteurs de sens ;
2. calculer les spécificités, et les étudier pour voir quels auteurs ont des tics très marqués ;
3. calculer la classification ;
4. calculer l'AFC.

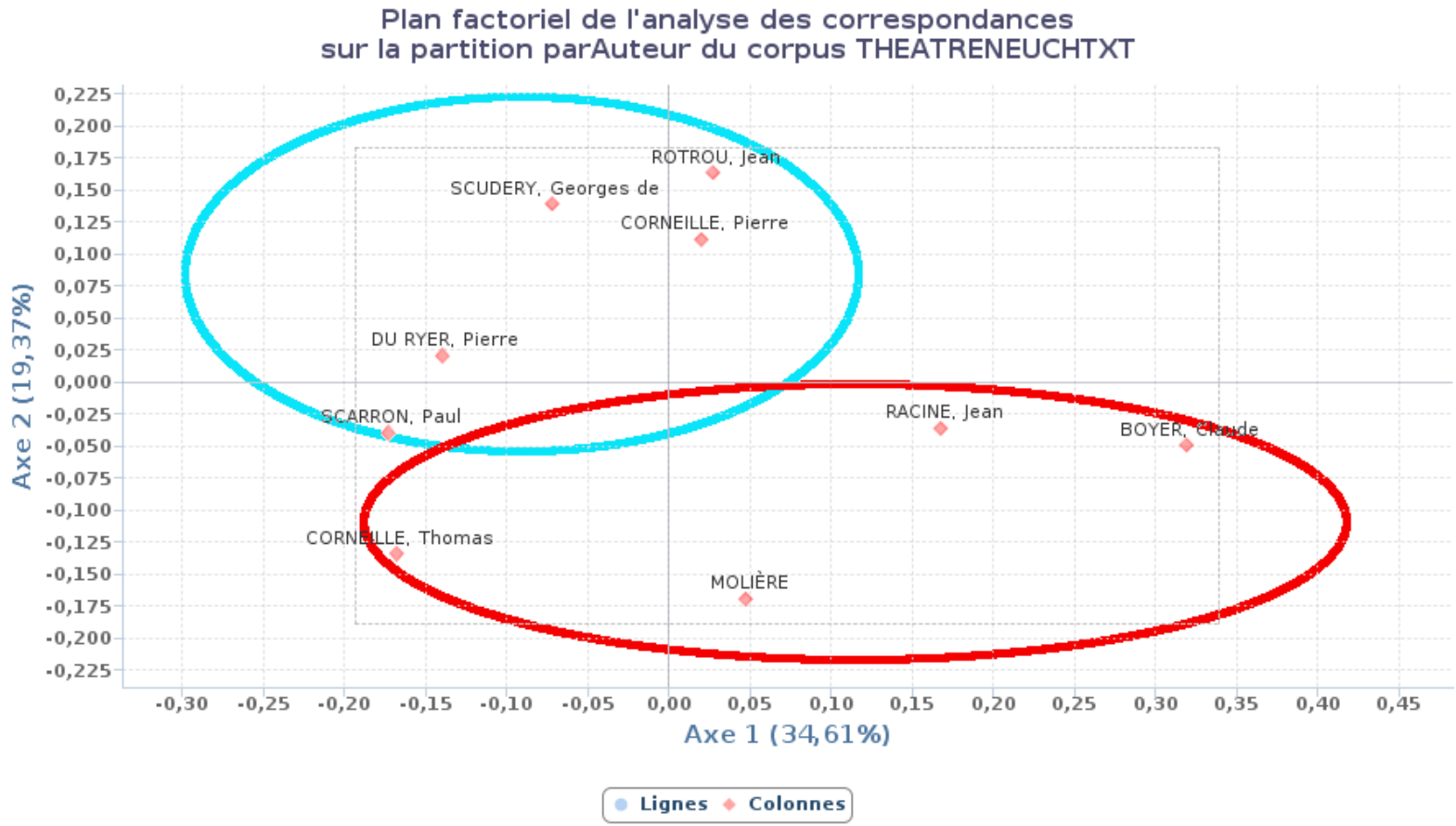
(Quelques) tics de langage de Claude Boyer



L'Analyse factorielle des correspondances



Une répartition surtout chronologique



Et maintenant : les textes eux-mêmes...

Sur la partition par textes,

1. Créer une table lexicale des 200 formes les plus fréquentes, retirer les mots porteurs de sens ;
2. calculer la classification ;
3. calculer l'AFC.

TO BE CONTINUED...

Avec R.

N.B. : ne pas oublier de télécharger la table de fréquence des formes par textes nettoyée, pour utilisation dans R.

Plan

Solutions

Interroger avec TXM : niveau 2, un peu de CQL

1. Modifier la requête pour permettre un mot, quel qu'il soit, entre seigneur et le pronom ;
   ```
   [word="seigneur"][][frpos="PRO.*"]
   [frlemma="être"]
   ```
2. la modifier, pour limiter aux cas où ce mot est une virgule ;
   ```
   [word="seigneur"][word=','][frpos="PRO.*"]
   [frlemma="être"]
   ```
3. l'éditer pour étendre aux cas où ce mot est une virgule OU un point d'interrogation.
   ```
   [word="seigneur"][word=',' |
   word='\?'][frpos="PRO.*"] [frlemma="être"]
   ```

Solutions

Interroger avec TXM : niveau 3, exercices

1. un mot contenant un chiffre.
   `[word=".*\d.*"]`
2. une forme commençant par 'a' et finissant par 'er' ;
   `[word="[a].*er"]`
3. une forme de deux lettres : une voyelle et 'h' ;
   `[word="[aeiouy]h"]`
4. 'seigneur' ou 'dame', suivi ou non d'un mot et suivi d'un pronom (n'importe quel type de pronom).
   `[word="seigneur" | word="dame"][]?[frpos="PRO.*"]`
5. la forme 'je', suivie d'entre 2 et 4 mots, et d'une virgule.
   `[word="je"][]2,4[word=","]`

Solutions

Interroger avec TXM : niveau 4, exercices

1. une forme contenant de la ponctuation ;
   `[word=".*\p{P}.*"]}`
2. un vers d'entre 4 et 5 mots ;
   `<l> []{4,5} </l>`
3. un vers débutant par un pronom personnel (`PRO:PER`) débutant par 't' ;
   `<l> [word="t.*" & frpos="PRO:PER"] []* </l>`
4. un vers se terminant par '-ron' (suivi ou non d'une consonne).
   `<l> [] * [word=".*ron[^aeiou]"] </l>`

Plan

Analyse par réduction des dimensions

- Des représentations graphiques simples permettent de résumer une grande partie de la relation entre deux variables.
- L'objectif des méthodes d'analyse de données que nous allons présenter est de remplir les mêmes offices dans le domaine plus délicat des statistiques multivariées.
- Ces analyses de données, qui visent à simplifier l'information pour la rendre lisible, sont appelées analyse par réduction des dimensions.

Analyse par réduction des dimensions : objectifs

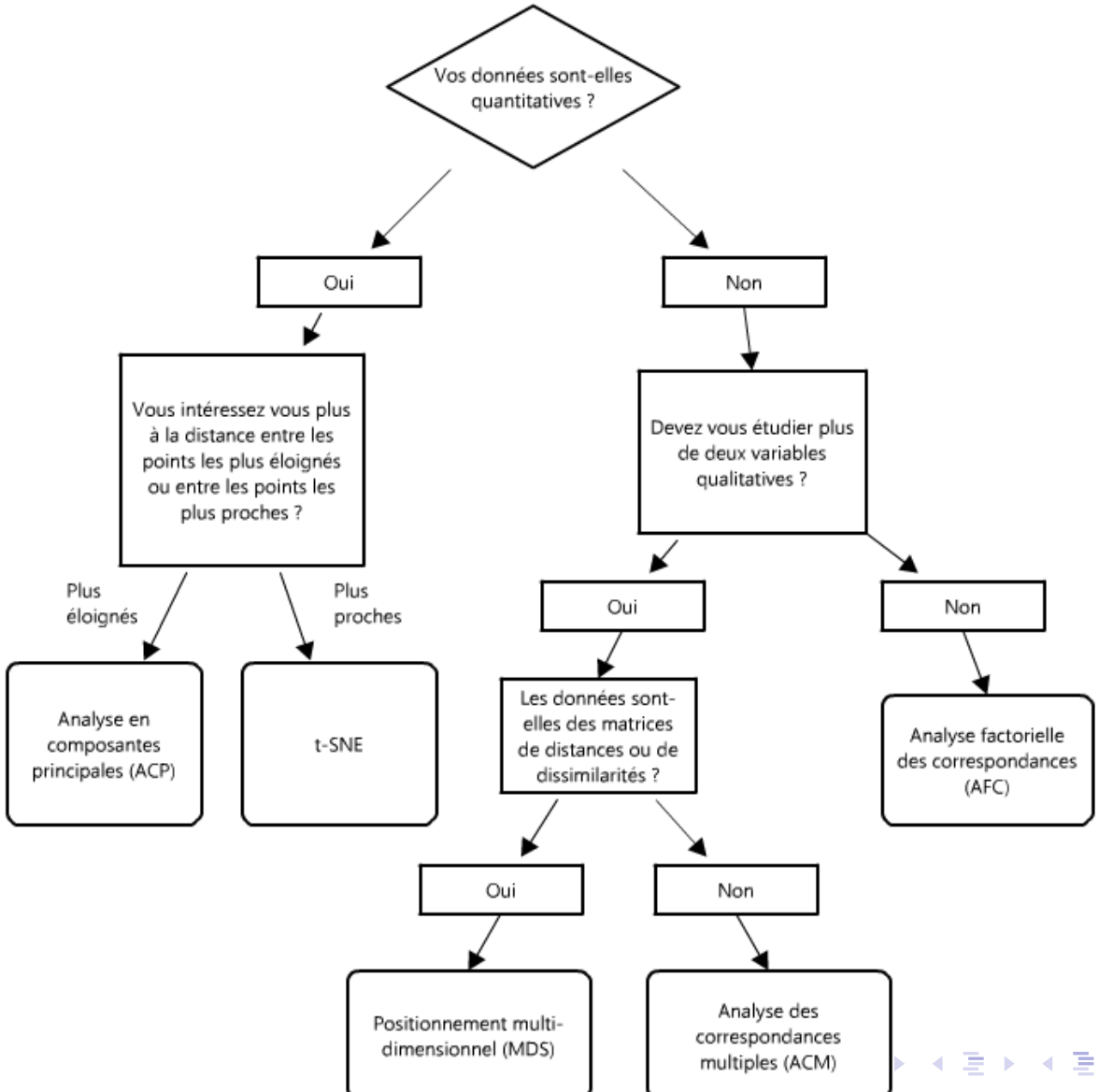
La valeur de ces méthodes est pour nous triple :

- **Visualiser** des données complexes pour y discerner des regroupements, des régularités, des typologies
- **Débruiter** les données, en supprimant de l'analyse les dimensions qui peuvent être considérées comme à négliger
- **Décorréler** : les axes créés au cours de ces analyses ne sont pas nécessairement des variables de la base, mais des construits n'ayant aucune corrélation entre eux.

Analyse par réduction des dimensions : méthode

- Dans une analyse factorielle, on cherche à déterminer les axes qui absorbent le plus d'inertie possible par rapport au centre de gravité du nuage de point.
- Concrètement, on va passer d'un nuage de points de grande dimension à un sous-espace de plus petite dimension, sur lequel on pourra réunir la plus grande quantité d'information possible.
- On choisit de déterminer des axes, définissant un plan sur lequel les points du nuage seront projetés.
- Ces axes sont choisis pour que les points projetés soient les plus dispersés possibles.

Analyse par réduction des dimensions : typologie



Analyse par réduction des dimensions : typologie

Trois de ces méthodes sont particulièrement usitées. Elles sont utiles dans des circonstances différentes :

1. Quand les variables sont quantitatives, on peut réaliser une Analyse en Composantes Principales (ACP).
2. Quand les variables sont qualitatives, on utilise des Analyses des Correspondances - la correspondance étant "l'équivalent" de la corrélation pour des variables qualitatives :
   - Quand les individus sont décrits par deux variables qualitatives, on peut construire un tableau de contingence et réaliser une Analyse Factorielle des Correspondances (AFC).
   - Quand les individus sont décrits par un jeu plus de deux variables qualitatives, on peut réaliser une Analyse des Correspondances Multiples (ACM).

Analyse par réduction des dimensions : ventilation

- Nécessité parfois de se débarrasser de modalités peu pertinentes, car elles brouillent le calcul, ou la visualisation
- Plutôt que la suppression, possibilité de ventiler : on remplace les modalités ne dépassant pas un effectif minimal par la valeur moyenne de l'échantillon sans ces modalités. On les neutralise ainsi.

Analyse factorielle des correspondances (AFC)

- L'AFC s'applique à des tableaux de contingence [1] c'est-à-dire des tableaux croisant deux variables qualitatives.
- L'AFC est en cela très distincte de l'ACP : les lignes et les colonnes jouent ici des rôles symétriques alors que la distinction entre lignes et colonnes, i.e. entre individus et variables, est majeure en ACP.
- Mais elle découle en fait de l'ACP : une AFC réalise une ACP sur le profil "ligne", une autre ACP sur le profil "colonne", et superpose les deux graphiques.

1. Voir Pearson K., "On the Theory of Contingency and Its Relation to Association and Normal Correlation", *Mathematical Contributions to the Theory of Evolution*, London : Dulau & Co, 1904

Classification ascendante hiérarchique

- La classification ascendante hiérarchique réalise un regroupement sous forme de dendogramme entre les différents individus d'un jeu de données.
- Elle est utilisable dans le cadre d'individus décrit par des **variables quantitatives**
- On peut toutefois tricher, en transformant des données qualitatives en données quantitatives - ce que nous verrons un peu plus tard.
  (La méthode la plus employée est de réaliser une **analyse factorielle** à partir des données quantitatives, et de se servir des coordonnées des points dans les axes factoriels pour réaliser la CAH.)

Métrique (1)

Pour réalisation ce type de classification, il faut d'abord définir comment on mesurera la "distance" entre deux textes. Il existe autant de possibilités qu'on le souhaite, mais certaines mesures sont plus usitées, et plus adaptées à nos types d'étude.

- **Distance euclidienne** :

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

C'est la mesure habituelle de la distance, celle du monde physique.

Métrique (2)

- **Distance de Manhattan**[2] :
  Somme des valeurs absolues des différences entre coordonnées.

$$\sum_{i=1}^{n} |x_i - y_i|$$

2. Pour en savoir plus sur cette géométrie particulière : Krause, E. F., *Taxicab Geometry : An Adventure in Non-Euclidean Geometry*. New York : Dover, 1986.

Métrique (3)

Elle pose en effet quelques problèmes, notamment pour des valeurs trop proches de 0, mais ce n'est pas un cas qui se présente dans nos usages en textométrie. Elle n'est pas implémentée dans le package `agnes`, mais se révèle pourtant souvent intéressante.

- **Distance de Canberra**[3] :
  Distance de Manhattan pondérée.

$$D_{Canb}(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

- **Utilité :** permet d'atténuer l'importance de certaines différences majeures d'un point de vue numérique. On s'intéresse plus à l'existence de différences qu'à l'importance quantitative de chacune de ces différences.

3. Lance, G. N. ; Williams, W. T. (1966). "Computer programs for hierarchical polythetic classification ("similarity analysis")". *Computer*

Métrique (4)

Le delta de Burrows, parfois dit “classique” (en stylométrie)[4].
Procédure un peu complexe de standardisation des comptes d’occurrences (z-transformation) et d’un changement de métrique (calcul de distance en utilisant la distance de Manhattan dont nous avons déjà parlé).
Considéré comme un outil aussi performant pour la prose que pour la poésie[5].

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - B_i}{\sigma_i}\right|$$

4. John Burrows, ‘Delta’ : a Measure of Stylistic Difference and a Guide to Likely Authorship, *Lit Linguist Computing* (2002) 17 (3).
5. David L. Hoover, “Testing Burrows’s Delta”, *Literary & Linguistic Computing* (2004) 19 (4).

Méthode

On doit ensuite choisir comment on regroupe les individus entre eux. Par défaut, le package propose d'utiliser la
Distance moyenne - "average" Calcule toutes les distances entre les différents points et en fait la moyenne
On peut aussi raisonner en terme d'extrémités
"Complete linkage" : calcule la distance maximale entre deux points
"Single linkage" : calcule la distance minimale entre deux points
L'algorithme le plus couramment utilisé est la :
Méthode de Ward : on calcule la distance entre les centres de gravité
Ces méthodes se rejoignent seulement dans des cas très spécifiques.

CAH avec R et cluster


```
agnes(x, diss = inherits(x, "dist"),
metric = "euclidean | manhattan",
stand = FALSE,
method = "average|single|complete|ward|weighted",
par.method,
keep.diss = n < 100, keep.data = !diss)
```

```
#importer la bibliotheque cluster
library(cluster)
#calculer la CAH
maCAH = agnes(CansosPondere100,
metric ="manhattan", method="ward")
#consulter les resultats
summary(maCAH)
#les tracer
plot(maCAH)
```

Plan

Bibliographie

Heiden, S., Magué, J-P., et Pincemin, B., « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », dans *Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, éd. Sergio Bolasco, Isabella Chiari, Luca Giuliano, Rome, 2010, t. 2, p. 1021-1032, https://halshs.archives-ouvertes.fr/halshs-00549779/fr/.

Pincemin, Bénédicte, Céline Guillot, Serge Heiden, Alexei Lavrentiev, et Christiane Marchello-Nizia, « Usages linguistiques de la textométrie : analyse qualitative de la consultation de la Base de Français Médiéval via le logiciel Weblex », *Syntaxe et Sémantique*, 9 (2008), p. 87–110, https://halshs.archives-ouvertes.fr/halshs-00355461.