

# TXM: la Textométrie à portée de clic

Crédits : Karën Fort (cours Plateformes pour le TAL)

---

Gaël Lejeune

14 novembre 2024

STIH/CERES, Sorbonne Université

# TXM : contexte

- Dans le cadre de l'ANR Textométrie (Heiden, Magué et Pincemin<sup>1</sup>)

- 
1. TXM : Une plateforme logicielle open-source pour la textométrie : conception et développement (Heiden *et al.* 2010)
  2. <http://textometrie.ens-lyon.fr>

# TXM : contexte

- Dans le cadre de l'ANR Textométrie (Heiden, Magué et Pincemin<sup>1</sup>)
- Analyse de grands corpus (structurés c'est encore mieux)
- Accepte différents formats : texte brut, XML, XML-TEI

- 
1. TXM : Une plateforme logicielle open-source pour la textométrie : conception et développement (Heiden *et al.* 2010)
  2. <http://textometrie.ens-lyon.fr>

# TXM : contexte

- Dans le cadre de l'ANR Textométrie (Heiden, Magué et Pincemin<sup>1</sup>)
- Analyse de grands corpus (structurés c'est encore mieux)
- Accepte différents formats : texte brut, XML, XML-TEI
- Multi-plateforme : Windows, Mac, Linux
- Disponible aussi via le web<sup>2</sup>

---

1. TXM : Une plateforme logicielle open-source pour la textométrie : conception et développement (Heiden *et al.* 2010

2. <http://textometrie.ens-lyon.fr>

# TXM : apports

- Open Source (pérennité)
- Passage à l'échelle (10 millions de mots)
- Intègre des outils externes (ex : Treetagger pour l'étiquetage)
- Combine de nombreux outils et visualisations
- Une communauté active :
  - Liste : <https://groupes.renater.fr/sympa//info/txm-users/>
  - Wiki : <https://groupes.renater.fr/wiki/txm-users/index>
  - Vidéos et tutoriels sur le site de TXM

# Fonctionnalités

## Analyses statistiques simples

- Index
- Concordances
- Cooccurrences

# Fonctionnalités

## Analyses statistiques simples

- Index
- Concordances
- Cooccurrences

## Analyses avancées

- Attirance contextuelle des mots et des expressions
- Spécificités lexicales
- Linéarité et organisation interne du texte
- Comparaisons de sous-corpus

# Fonctionnalités

## Analyses statistiques simples

- Index
- Concordances
- Cooccurrences

## Analyses avancées

- Attirance contextuelle des mots et des expressions
- Spécificités lexicales
- Linéarité et organisation interne du texte
- Comparaisons de sous-corpus

Principe de base en textométrie (ma traduction) :  
l'outil dégrossit, l'humain interprète

## Quelques usages

---

# Lexique et Index

- Liste de formes (ou de tokens)
- Fréquence d'apparition
- Lemmatisation/étiquetage (Treetagger) :
  - Forme dictionnaires (suppose corpus monolingue pour TXM)
  - NOM, ADJ, VER, ADV + morphologie

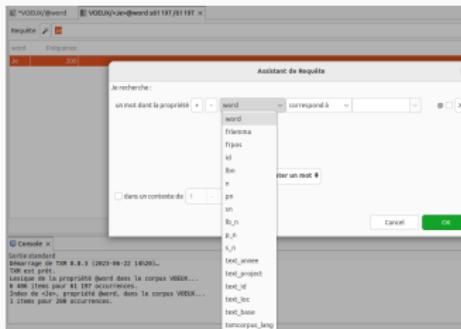
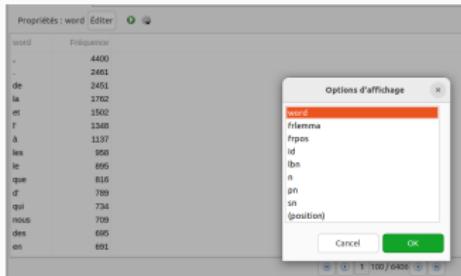
# Lexique et Index

- Liste de formes (ou de tokens)
- Fréquence d'apparition
- Lemmatisation/étiquetage (Treetagger) :
  - Forme dictionnaires (suppose corpus monolingue pour TXM)
  - NOM, ADJ, VER, ADV + morphologie
  - **Étiquetage/lemmatisation : tâches maîtrisées mais non résolues**

# Lexique et Index

- Liste de formes (ou de tokens)
- Fréquence d'apparition
- Lemmatisation/étiquetage (Treetagger) :
  - Forme dictionnairiques (suppose corpus monolingue pour TXM)
  - NOM, ADJ, VER, ADV + morphologie
  - **Étiquetage/lemmatisation : tâches maîtrisées mais non résolues**
- Revoir le mot dans son contexte
- Allers et retours entre le lexique et le corpus

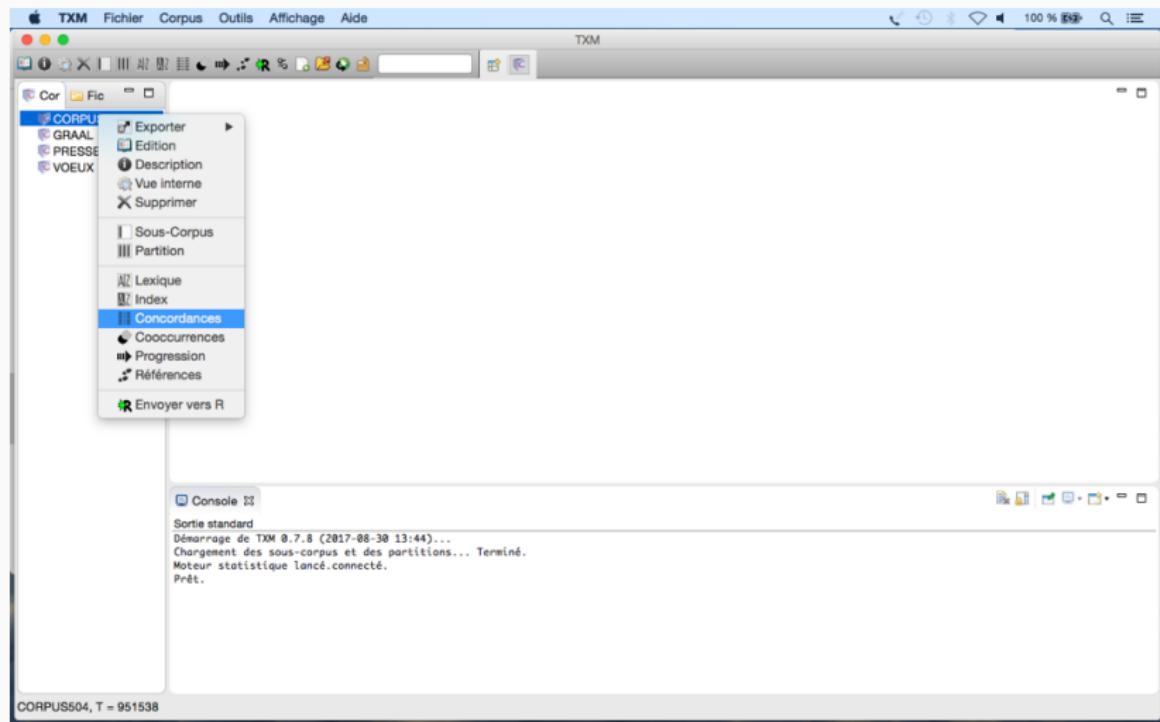
# Lexique VS Index



# Concordances

- Observation synthétique des occurrences d'une forme (d'un motif) :
  - ses contextes d'apparition
  - triés de différentes façons
- Utilisations :
  - distribution dans le corpus
  - expressions dérivées
  - structures grammaticales

# Concordances (I)



# Concordances (II)

Requête : **effet** Pivot: word Editer Chercher

Cles de tri : #1 Aucun #2 Aucun #3 Aucun #4 Aucun Tri

1 - 100 / 377 Cacher paramètres

text_id	Contexte gauche	Pivot	Contexte droit
CRurgence	avérera non répondre le DV étant également sans	effet	. Finalement une laparotomie exploratrice à la recherche d'un hypothétique foyer
CRurgence	, contusion de la pointe temporelle gauche sans	effet	de masse ni oedème. TDM thoracique : hématome extrapleural gauche limité
CRurgence	objectivera une majoration de la contusion temporelle avec	effet	de masse sur le VL mais sans déplacement des structures médianes ni
CRurgence	cardiogénique associée. une insuffisance respiratoire aiguë avec	effet	shunt important, liée à la pneumopathie à Légionnelle et à un
artPress	de nous, les indépendantistes québécois sont en	effet	favorables à tout ce qui permet au Québec de s'ouvrir au
artPress	le Québec soit libre c'est, en	effet	, ce dont il s'agit. Cela aboutira forcément, à
artPress	républicain fera appel ; il se dit en	effet	plus que satisfait de la décision : " Qui ne serait pas
artPress	mardi 24 novembre, ces chercheurs ont en	effet	épinglé un passage du document qui dit ceci : " Comment éviter
artPress	Mais elle est plus que bancale. En	effet	, il y a un hic : Georges Frêche a précisément été
artPress	et son associé britannique Clive Bowen veulent en	effet	construire " le premier circuit de F1 durable ". Article suivi
artPress	espérer mieux. Les électeurs potentiels devaient en	effet	imprimer eux-mêmes leurs bulletins. Ecologie et pauvreté obligent. A Paris
artPress	. La lutte contre l'abstention était en	effet	au cœur du programme de Cannabis sans frontières, qui espérait offrir
artPress	: 1. Achèver une clarification séculaire En	effet	, la social-démocratie dont vous êtes les héritiers entretient avec l'idéal
artPress	? 2. Achèver une clarification contemporaine En	effet	, votre parti ayant abondamment participé à l'entassement de lois sécuritaires
artPress	. 3. Achèver le parti : en	effet	, la référence de Valls est manifestement transalpine. Même fascination pour le
artPress	Toute la problématique du téléphone portable est l'	effet	à long terme des faibles doses. " Depuis, plusieurs procédures
artPress	qui pourraient également toucher sa hiérarchie. En	effet	, deux courriers adressés à Hervé Morin, en 2007 et 2009
artPress	évitée, si le lien de cause à	effet	entre les tirs et l'incendie était prouvé : " C'est

Console

Sortie standard

Démarrage de TMX 0.7.8 (2017-08-30 13:44)...

Chargement des sous-corpus et des partitions... Terminé.

Moteur statistique lancé.connecté.

Prêt.

Concordance de <"effet"> dans le corpus CORPUSS504

377 occurrences

## Concordances : le pivot

Le pivot peut être :

- **mot**, séquence de mots

## Concordances : le pivot

Le pivot peut être :

- **mot**, séquence de mots
- **lemme**, une séquence de lemmes

## Concordances : le pivot

Le pivot peut être :

- **mot**, séquence de mots
- **lemme**, une séquence de lemmes
- **partie de mot(s)** (expression régulière)
- **motif** en langage CQL (Corpus Query Language)

# Concordances : le pivot

Le pivot peut être :

- **mot**, séquence de mots
  - **lemme**, une séquence de lemmes
  - **partie de mot(s)** (expression régulière)
  - **motif** en langage CQL (Corpus Query Language)
  - → attention, pas très intuitif!
  - mais de la richesse : motifs syntaxiques, lexicaux, lexico-syntactiques
- ...

# Propriétés du pivot

\*VOEUX peut

Requête peut

Paramètres

Clés de tri #1 Contexte ; #2 Aucun ; #3 Aucun ; #4 Aucun ; Tri

Références

Vue	text_id ; text_annee ;
Tri	text_id ...
Taille	

Options d'affichage

frlemma	word
id	frpos
lbn	
n	
pn	
sn	

Pivot

Pivot	frlemma	Contexte droit
peut_VER:pres	pouvoir	et doit être, pour le bien des hommes, l'
peut_VER:pres	pouvoir	l'aider, nous offrons d'avance notre con
peut_VER:pres	pouvoir	l'oublier ? - en raison de certaines évide

Cancel OK

qui monte. Cela eut lieu - qui

# Fréquences lexicales

Observer **types** et **tokens**

Table des fréquences : distribution par type (Mots, Lemmes, POS ...)

Assez classiquement (Zipf) on trouve :

- Les plus fréquents : des mots grammaticaux

---

3. Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle, thèse de l'Université de Caen

# Fréquences lexicales

Observer **types** et **tokens**

Table des fréquences : distribution par type (Mots, Lemmes, POS ...)

Assez classiquement (Zipf) on trouve :

- Les plus fréquents : des mots grammaticaux
- Ensuite des mots du thèmes / du genre textuel (si corpus homogène)
- Sur le corpus 504 :
  - 0,0002% des formes produisent plus de 20% des occurrences
  - Longue traîne : les hapax représentent 50% du vocabulaire (18 000)
  - Les mots rares sont très fréquents (Lardilleux 2010<sup>3</sup>)

---

3. Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle, thèse de l'Université de Caen

# Lexique VS Index : 2 fonctionnalités différentes

## Lexique

- Calcule la fréquence pour une propriété de mot donnée :
- forme, lemme ...
- mais pas d'expressions complexes
- Première visualisation du corpus : thèmes, hapax

## Index

- Calcule la fréquence d'une expression (mot unique ou non)
- agit comme un filtre sur le lexique
- adapté à la recherche à tâtons dans le corpus

# Lexique (I)

Cor Fic

▼ CORPUS504  
[frlemma="effet"]  
GRAAL  
PRESSEPIAPER1  
VOEUX  
▼ VOLCANSIENTFR  
word

Propriétés : word

Seuils : Fmin : 1  Fmax : 9999999  Vmax : 9999999  Résultats par page : 100

<  < 1 - 100 / 25299 >  >|

t 785115 , v 25299 , fmin 2 , fmax 40096

word	Fréquence
,	40096
de	39168
.	30754
la	16890
et	16541
des	14458
)	13302
(	13074
à	12468
les	10600
du	9969
le	9788
l'	9446

# Index (I)

Requête :  Propriétés : word

Seuils : Fmin : 1 Fmax : 9999999 Vmax : 9999999 Résultats par page : 100

[<] [<] 1 -39 / 39 [>] [>]

t 4716 , v 39 , fmin 2 , fmax 1546

word	Fréquence
volcan	1546
volcanique	844
volcaniques	780
volcans	714
volcanisme	478
volcanologique	70
volcanologiques	42
volcanites	38
volcanologie	26
volcano-tectonique	26
volcanologues	18
volcan-bouclier	16
volcanic	16
volcano-sédimentaires	16
volcano-sédimentaire	10
volcano	8
volcanoclastiques	6
volcanologue	6

Console

Sortie standard

449 occurrences

Index de <input type="checkbox"/> avec la propriété [word] dans le corpus VOLCANSIENTFR  
Terminé : 25299 items pour 785115 occurrences.

Index de <input type="checkbox"/> avec la propriété [word] dans le corpus VOLCANSIENTFR  
Terminé : 39 items pour 4716 occurrences.

## Exemples d'études sur corpus

---

## Analyse de discours politique (avant TXM)

- **Etude chronologique du discours syndical (Salem 1993)**
- Vocabulaire présidentiel : le cas de F. Mitterrand (Labbé 1990)
- Richesse lexicale des discours politiques (Véronis 2007)

## Étude chronologique

- Objectif : dégager des évolutions dans les usages lexicaux
- Corpus : textes de congrès syndicaux (1973-1988)
- Moyens : analyse des termes "salariés" et "travailleurs"

# Discours syndicaux (I)

## Annexe

Tableau B : Les fréquences relatives des formes *salariés* et *travailleurs* dans les six périodes du corpus CFDT

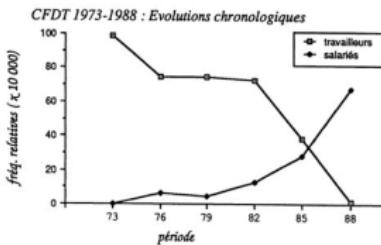
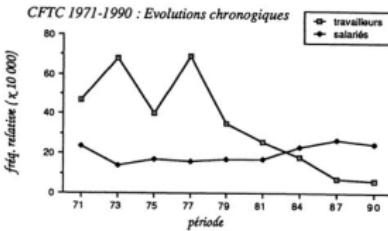


Tableau C: Les fréquences relatives des formes *salariés* et *travailleurs* dans les neuf périodes du corpus CFTC



## Discours syndicaux (II)

<b>Contexte terme 1</b>	<b>Eff.</b>	<b>Contexte terme 2</b>	<b>Eff.</b>
tous les travailleurs	22	tous les salariés	6
ensemble des travailleurs	19	ensemble des salariés	4
pour les travailleurs	13	pour les salariés	8
catégories de travailleurs	6	catégories de salariés	6
intérêts des travailleurs	6	intérêts des salariés	3
aspirations des travailleurs	6	aspirations des salariés	3
permettre aux travailleurs	6	permettre aux salariés	2
expression des travailleurs	2	expression des salariés	3

**Table 1** – Contextes d'apparitions des deux termes (Salem 1993) remis par ordre décroissant

## II. Étude du vocabulaire présidentiel, le cas de F. Mitterrand

- Objectifs : étudier la spécificité du vocabulaire présidentiel, cad, à quelle fréquence certains mots sont-ils employés, et dans quels contextes ?<sup>4</sup>
- Corpus : interventions radio-télévisées de F. Mitterrand, 1981-1988
  - 68 interventions
  - 305124 mots
  - Environ 40h de diffusion

---

4. Labbé, D. (1990). *Le vocabulaire de François Mitterrand*. presses de la fondation nationale des sciences politiques, Paris, France.

# Étude du vocabulaire présidentiel le cas de F. Mitterrand

- Fréquence des mots

Quel est le vocabulaire propre à ce président ?  
Comparaison avec d'autres corpus

Vocables	Fréquence
acquis n m	16
acquis adj	12
acquis	10
acquise	1
acquiesces	1

Tableau 2. Extrait de la table des formes et des lemmes du corpus Mitterrand  
(Labbé, 1990)

# Étude du vocabulaire présidentiel le cas de F. Mitterrand

- Tableau : lemmes / formes : acquis / acquise
- **Résultats :**
  - 20 substantifs les plus fréquents : français, pays, homme, ... substantifs que l'on retrouve chez d'autres hommes politiques (Chirac, De Gaulle). Pas de marquage idéologique
  - Utilisation des verbes plus remarquable :
    - présence de verbes désignant plutôt la pensée que l'action
    - un certain déficit en verbes exprimant la connaissance (savoir, connaître)
    - utilisation importante des modalités pouvoir, vouloir, devoir.

# Étude du vocabulaire présidentiel le cas de F. Mitterrand

- Les contextes d'utilisation des mots  
Le locuteur fait-il un usage particulier des mots ?
- Ex : utilisation fréquente du pronom personnel je
  - Plutôt banal dans le discours
  - À quels autres mots est-il fortement associés ?
  - Ou à l'inverse, avec quels mots n'apparaît-il pas ?

# Étude du vocabulaire présidentiel le cas de F. Mitterrand

- Je est fortement associé aux mots suivants :
  - **Verbes** : croire, dire, penser, répéter, souhaiter, répondre, espérer
  - **Noms** : ministre, Français, président
  - **Adjectifs** : heureux, favorable, sûr, partisan
- Je a tendance à exclure :
  - Pouvoir, falloir, permettre, exister
  - Plan, chômage, entreprise, exemple, problème

## Étude du vocabulaire présidentiel le cas de F. Mitterrand

- Je + Verbes de parole, de pensée, de volonté
- Je - Verbes marquant la possibilité ou la nécessité

# Étude du vocabulaire présidentiel le cas de F. Mitterrand

- Je + Verbes de parole, de pensée, de volonté
- Je - Verbes marquant la possibilité ou la nécessité
- Observations → interprétations.

### III. Richesse lexicale des discours politiques

Blog : "Technologie du Langage" <sup>5</sup>

---

5. Véronis, J. (2007) "Texte : richesse lexicale ". Blog Technologie du Langage, 3 mars 2007. <http://blog.veronis.fr/2007/03/texte-richesse-lexicale.html>

# Richesse lexicale des discours politiques

- "On me pose souvent la question : qui de nos candidats a le vocabulaire le plus riche ? (...) la réponse est tout sauf simple."
  - "Comment quantifier la richesse lexicale d'un text de façon rigoureuse ?"
- 
- Idée : trouver un indice

## Exemple

Dans le texte étudié, compter :

- le nombre total de mots du texte
- le nombre de mots différents
- faire le rapport entre les deux

# Richesse lexicale des discours politiques

## Exemple

- Ségolène Royal, Villepinte :

- 12819 mots
- 2707 mots différents
- Rapport = 0.21

"Pour clarifier les choses, on parle d'occurrences et de formes : 12819 occurrences, 2707 formes".

- Ségolène Royal, voeux du 4 janvier :

- 1119 formes
- 3483 occurrences
- rapport = 0.32

→ Cela signifie-t-il que son discours de voeux était plus riche que son discours à Villepinte ?

## Richesse lexicale des discours politiques

- On ne peut rien conclure car :
  - Textes de tailles différentes
  - Textes courts : ont tendance à avoir un rapport formes / occurrences plus élevé que les textes longs...
- Indice utile si les textes sont de tailles très voisines, sinon inexploitable
- Si grand nombre de textes :
  - pour chaque texte, reporter le nombre d'occurrences et le nombre de formes sur un graphique
  - Exemple : Discours 2007

## Richesse lexicale des discours politiques

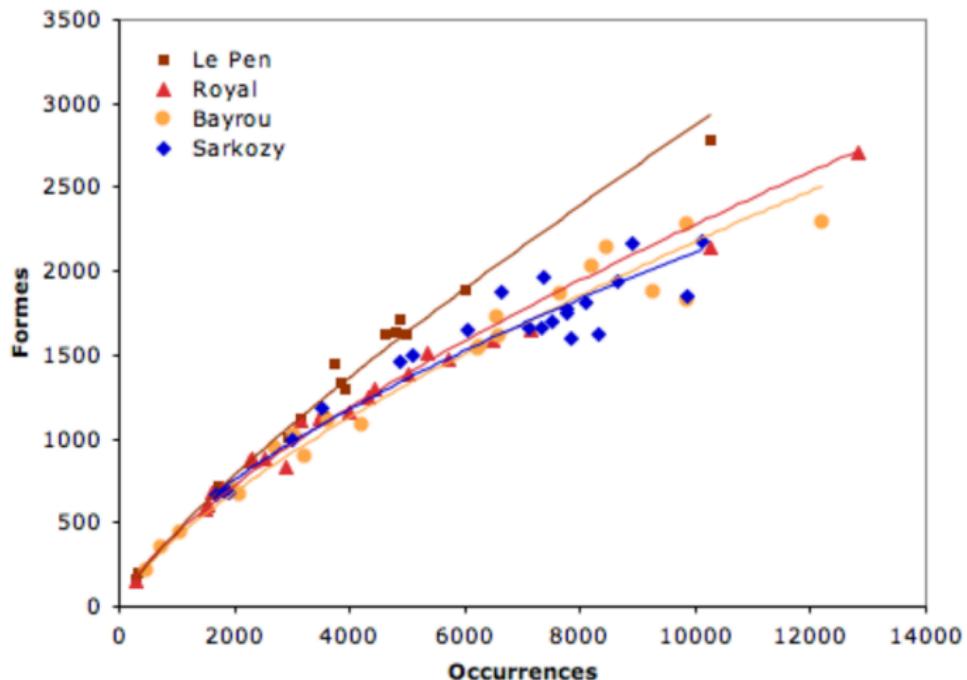


Figure 2. Nombre d'occurrences et de formes par texte, par candidat.

[Nb : 1 point = 1 discours]

# Richesse lexicale des discours politiques

- Courbe de tendance : montre des différences entre les autres
- COurbe de Le Pen au-dessus des autres
- Les discours de Le Pen sont plus "riches" lexicalement que ceux des autres candidats.
- "Riches" :
  - n'implique pas de jugement de valeur, ni de compréhensibilité
  - Davantage de mots différents, vocabulaire plus varié.

## TXM en pratique

---

## Importer un corpus

- Importer, presse-papier OK
- des fichiers texte brut OK

## Importer un corpus

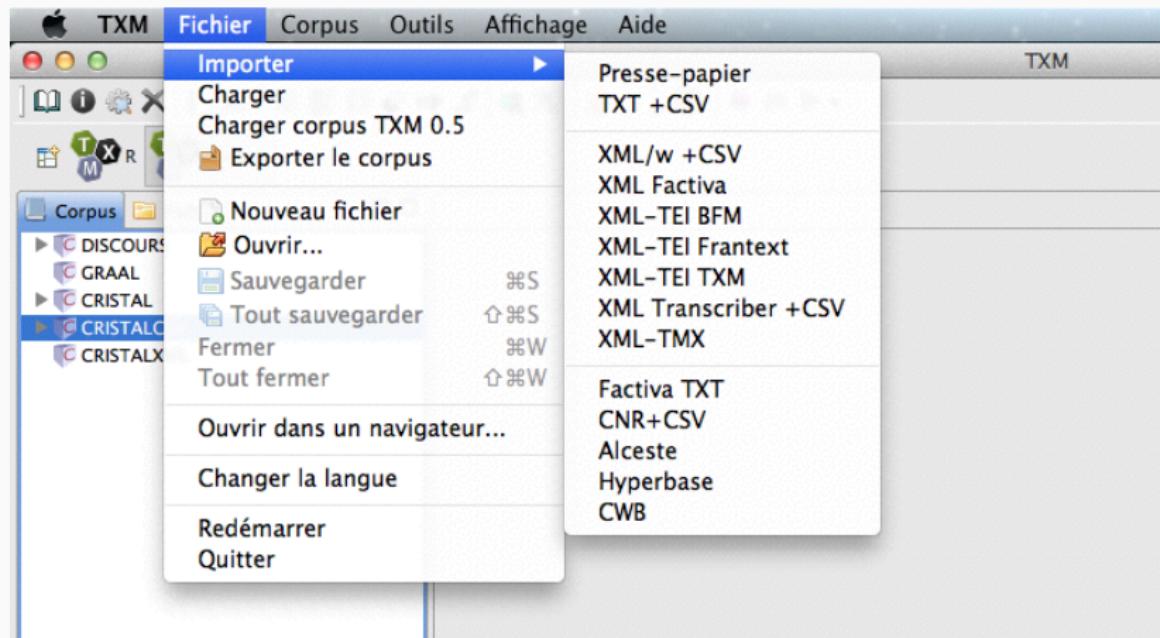
- Importer, presse-papier OK
- des fichiers texte brut OK

**Mais on peut avoir des formats plus riches :**

- XML (avec méta-données explicites donc)
- Formats d'autres logiciels (Hyperbase, Alceste)
- XML-TEI (Frantext, Transcriber ...)

NB : l'installation/l'usage de TREE TAGGER n'est pas obligatoire, cela augmente juste les possibilités de requêteage

# Importer : Formats



# Importer : Préparation des données

Import XML/w + CSV de TEST

## Paramètres d'import du module XML/w + CSV

- [Sélectionner le répertoire des fichiers sources](#)
- Régler les paramètres d'import dans les sections ci-dessous.
- [Lancer l'import du corpus.](#)

▼ Description du corpus TEST (/home/rundimeco/Documents/enseignement/2019\_IngenierieLangues/)

Nom complet, auteur, date, licence, commentaire...

rundimeco  
17 mars 2020, 07h57

▼ Langue principale

[Annoter le corpus](#)

[Deviner :](#)

[Sélectionner :](#) fr ▼

► Segmentation lexicale

# Tag or Not to tag ?



## Formats : mais pourquoi ?

- Un format "libre" favorise l'émetteur
- ... mais constraint le récepteur

---

6. De Busser, 2006 Information extraction and information technology, pages 1–22.  
Springer, Berlin, Heidelberg

## Formats : mais pourquoi ?

- Un format "libre" favorise l'émetteur
- ... mais constraint le récepteur
- Un format rationnel (structuré) c'est le contraire

"Texts in natural language are not unstructured, they are computationally opaque<sup>6</sup>

---

6. De Busser, 2006 Information extraction and information technology, pages 1–22.  
Springer, Berlin, Heidelberg

## Formats : mais pourquoi ?

- Structurer c'est expliciter pour la machine
- C'est permettre d'exploiter la dynamique interne du corpus

---

7. Rastier 2002, "Enjeux épistémologiques de la linguistique de corpus"

## Formats : mais pourquoi ?

- Structurer c'est expliciter pour la machine
- C'est permettre d'exploiter la dynamique interne du corpus
- "Le texte est pour une linguistique évoluée l'unité minimale<sup>7</sup>"

Constraining but a major interest of TXM are the advanced functions that make use of sub-corpora (thus meta-data)

---

7. Rastier 2002, "Enjeux épistémologiques de la linguistique de corpus"

# Exemple de problème de Structuration

## Documents sauvegardés • 50 documents

Le Figaro

27 avril 2011

### La « digital mum », nouvel eldorado des marques

COMMUNICATION Le profil de la « digital mum » s'affine. Avec la généralisation d'Internet et des nouvelles technologies, elle est même en passe d'éclipser « la ménagère de moins de 50 ans ...

8

Le Figaro

27 juin 2011

### La « digital mum » part en vacances avec Internet

... plus, elles resteront connectées tout l'été. Le premier observatoire WebMediaGroup/KR Media sur les « digital mums » est sans appel : ces mères au foyer avec enfant, qui se connectent à Internet au ...

10

Le Figaro

25 octobre 2011

### La Digital Mum intègre Médiamétrie

... elle est identifiée et désormais elle sera prise en compte par les régies publicitaires. La Digital Mum, cette femme active qui surfe régulièrement sur le Web, découverte il y a un peu ...

12

Structuration visuelle : évidente pour l'humain, opaque pour la machine  
(*not machine readable*)

# Exemple de problème de Structuration

## LE FIGARO

Nom de la source

Le Figaro

Type de source

Presse • Journaux

Périodicité

Quotidien

Couverture géographique

Nationale

Provenance

France

p. 26



Mardi 25 octobre 2011

Le Figaro • no. 20910 • p. 26 • 311 mots

## La Digital Mum intègre Médiamétrie

La nouvelle ménagère pourra être prise en compte par les régies publicitaires.

*Paule Gonzalès*

**I**NTERNET Elle existe, elle est identifiée et désormais elle sera prise en compte par les régies publicitaires. La Digital Mum, cette femme active qui surfe régulièrement sur le Web, découverte il y a un peu moins d'un an par WebMediaGroup et développée par l'agence médias KR Médias, fait son entrée chez Médiamétrie, le temple de la mesure d'audience.

*« Si elle n'est pas encore une cible de*

ernes aussi bien pour réserver les vacances sur le Web que pour effectuer les achats de Noël.

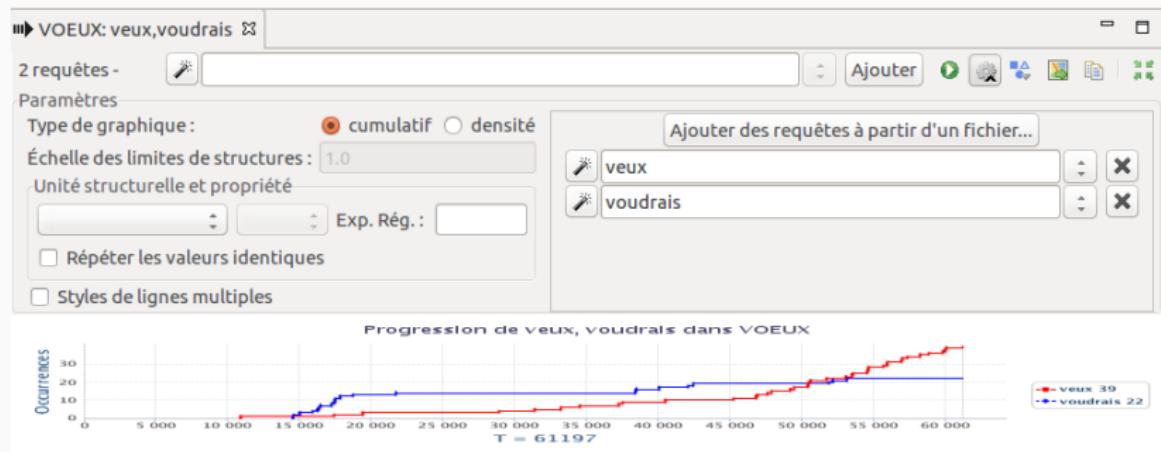
### Active pour les fêtes de fin d'année

Une nouvelle étude constate que « 70 % des Digital Mums utiliseront Internet pour leurs achats de fin d'année et pour 19 % d'entre elles c'est même précisément durant cette période de l'année qu'elles prévoient d'acheter le plus sur la Toile ». Toutefois, si la Digital Mum

# Exemple de donnée structurée

```
▼<corpus>
  <article titre="La « digital mum », nouvel eldorado des marques Un baromètre trimestriel créé par KR Media et WebMediaGroup permettra de mieux comprendre ses comportements." date="2011 04 27" journal="Le Figaro, no. 20756">COMMUNICATION Le profil de la « digital mum » s'affine. Avec la généralisation d'Internet et des nouvelles technologies, elle est même en passe d'éclipser « la ménagère de moins de 50 ans », cible commerciale très convoitée née dans les années 1960. Pour la première fois, l'impact des nouveaux médias dans la consommation et dans la vie des mères de famille est étudié. L'agence KR Media, conseil en stratégie et achat d'espaces publicitaires, ne s'y trompe pas. Elle propose à ses clients « une vision de la ménagère de moins de 50 ans qui soit plus en phase avec la réalité de notre société. La »digital mum* relie parfaitement les mondes physiques et numériques dans lesquels nos annonceurs déplacent leurs actions marketing. La »digital mum* est ainsi la nouvelle cible universelle que nous devons mieux comprendre. » Aussi vient-elle de signer avec WebMediaGroup, inventeur de la « digital mum », un partenariat pour mieux la définir et la suivre dans ses comportements médias et d'achat, via un baromètre trimestriel. A terme, ce baromètre pourrait séduire d'autres acteurs dont Médiamétrie qui a du mal à vendre à l'international le concept de ménagère de moins de 50 ans. La « digital mum » est « une femme ayant au moins un enfant à charge et se connectant au moins une fois par semaine à Internet », explique Isabelle Bordry, PDG de WebMediaGroup. En France, les « digital mums » représentent 17 % de la population des 15 ans et plus, soit 8,7 millions. C'est presqu'autant que les ménagères de moins de 50 ans, qui sont 10,7 millions en France. D'ailleurs, elles se confondent un peu. Ainsi 80 % des « digital mums » sont des ménagères de moins de 50 ans. Selon Isabelle Bordry, « la »digital mum* a en moyenne 40 ans mais le sentiment d'en avoir 33 et déclare agir autant par intuition que par raison ». Enfin, 45 % de ces dernières ont un revenu mensuel par foyer supérieur à 2 700 euros net contre 39 % pour la ménagère de moins de 50 ans. Cible non homogène L'objectif de cette étude est de mieux cerner les caractéristiques de cette nouvelle génération de consommatrices. Cela « devrait sensibiliser les marques non seulement à l'évolution de leur communication, mais aussi à celle des services et
```

# Progression (I)



# Progression (II)

TXM

Fichier Édition Corpus Outils Affichage Aide

Corpus Fichier ANNEE1999 PRES VOEUX

président [frpos = "VER:cond"] Je: word

2 requêtes - Ajouter

Paramètres

de graphique :  cumulatif  densité

Nombre des limites de structures : 1.0

Structure et propriété

text loc Exp. Rég. :

Répéter les valeurs identiques

Types de lignes multiples

Ajouter des requêtes à partir d'un fichier...

[frpos = "VER:cond"]

[frpos = "VER:pres"]

Progression de [frpos = "VER:cond"], [frpos = "VER:pres"] dans VOEUX  
(structure : text, propriété : loc, REGEX de filtrage : )

F2

T = 61197

frpos = "VER:cond" 103

frpos = "VER:pres" 3782

Console

Sortie standard

103, 3782 position(s.)

## Partitions et sous-corpus

- Sous-corpus : regroupement "minimal" déterminé selon les métadonnées
- Partition : un ensemble de sous-corpus
- On peut ensuite "opposer" des partitions pour faire émerger des phénomènes par contraste

## Spécificités lexicales

- Que signifie une fréquence absolue ?
- Comment comparer deux corpus ?
- La fréquence relative est un premier outil

## Spécificités lexicales

- Que signifie une fréquence absolue ?
- Comment comparer deux corpus ?
- La fréquence relative est un premier outil
- La spécificité lexicale : est-ce que la fréquence du mot est étonnante par rapport à la probabilité attendue ?

## Cooccurrences

- basées sur le calcul de spécificité
- "Attriance" (VS répulsion) statistique des mots
- Probabilité d'être voisins théorique (*prior*)
- Probabilité observée (*observation*)
- Si  $observation >> prior$  alors c'est remarquable