



Outils informatiques d'analyse linguistique - TXM présentation et commandes de base

Karën Fort

karen.fort@paris-sorbonne.fr / <http://karenfort.org>

17 novembre 2016



Quelques sources d'inspiration

par ordre d'importance décroissant

- Atelier TXM du 25 et 26 septembre 2014
- Manuel de TXM :
<http://txm.sourceforge.net/doc/manual/manual1.xhtml>
- Vidéo : http://txm.sourceforge.net/enregistrement_atelier_initiation_TXM_fr.html
- B. Pincemin (IHRIM) et S. Heiden (IHRIM)

- 1 Sources
- 2 Introduction**
 - Présentation
 - Premiers pas avec le corpus Vœux
- 3 Compter et voir en contexte
- 4 Pour finir
- 5 Bibliographie

Une communauté active

- formations
 - liste de diffusion active
 - communauté d'utilisateurs et de développeurs
 - ▶ logiciel libre
 - ▶ supportant Unicode
 - ▶ multi-plateformes (Java)
 - ▶ modulaire (R et CQP)
 - documentation sous différentes formes (pdf, vidéo, pages Web)
- + version portail Web :
- <http://portal.textometrie.org/demo/?locale=fr>



La documentation et la vitalité de la communauté sont des critères fondamentaux dans le choix d'un logiciel

Quelle communauté ?

Sciences humaines et sociales :

- archives historiques
- dépouillement d'enquêtes avec questions ouvertes
- œuvres littéraires
- corpus scientifiques
- etc.

Textométrie ?

Spécificité française

Historiquement, évolution et élargissement avec les avancées techniques (annotations, structuration) :

- **lexicométrie** : statistiques lexicales (sur les mots)
- **logométrie** : statistiques sur les discours
- **textométrie** : statistiques sur les textes

→ les calculs sont délégués à l'ordinateur, mais le chercheur reste maître de l'interprétation

Particularités de TXM

- **interface** très complète
- **robustesse** : permet de traiter jusqu'à 10 millions de mots
- **puissance** : permet d'intégrer toutes sortes de traitement *via* le logiciel R (de statistiques)
- **rapidité** : permet d'interroger des millions de mots très efficacement *via* CQP (Corpus Query Processor)

TXM

TXM permet d'explorer les corpus et de les analyser manuellement

⇒ outil d'analyse très pratique (indispensable ?)

Téléchargement et interface

Préalable

Télécharger le corpus Vœux :

```
http://sourceforge.net/projects/txm/files/corpora/voeux/  
voeux-bin.txm/download
```

Différents espaces, à explorer :

- onglets
- menuS (3 modes d'accès)
- console

Commandes de bas niveau

- charger (un corpus déjà importé auparavant)
- édition
- description

Corpus Vœux

Chargez le corpus Vœux

Que pouvez-vous dire sur le corpus Vœux grâce à TXM ?

Que manque-t-il ?

Commandes de bas niveau

- charger (un corpus déjà importé auparavant)
- édition
- description

Corpus Vœux

Chargez le corpus Vœux

Que pouvez-vous dire sur le corpus Vœux grâce à TXM ?

Que manque-t-il ?

→ le nombre entre parenthèses après `id` sous `text` donne le nombre de textes

→ mais il manque la licence et un descriptif !

Charger vs importer

Charger : corpus déjà importés dans TXM auparavant

Importer : corpus brut (txt, XML, voire en provenance du presse-papier)

Import *via* le presse-papier

- aller sur le site Web
- copier le contenu de la page (CTRL+C)
- dans TXM, sélectionner Fichier/Importer/Presse-papier
- tada !

Réglages

- vue interne
- ajout d'informations
- changement d'affichage

- 1 Sources
- 2 Introduction
- 3 Compter et voir en contexte**
 - Lexique
 - Concordance
 - Index et cooccurrences
- 4 Pour finir
- 5 Bibliographie

Qu'est-ce que le lexique pour TXM ?

- liste de formes (par défaut, mais paramétrable)
- fréquences d'apparition
- lemmatisation et étiquetage (par défaut) avec TreeTagger [Schmid, 1997], mais possibilité d'importer des corpus pré-annotés
- [lien](#) vers la concordance



Le contexte est fondamental dans TXM (seule la remise en contexte permet l'analyse)

Qualité de l'étiquetage morpho-syntactique

ou *POS tagging*

Exactitude (*accuracy* en anglais, à ne pas confondre avec la précision) :

- TreeTagger (1994) : 95,7 % [Allauzen and Bonneau-Maynard, 2008]
- ME1t (2010) : près de 98 % [Denis and Sagot, 2010]



Quelle différence concrète ?

Qualité de l'étiquetage morpho-syntaxique

ou *POS tagging*

Exactitude (*accuracy* en anglais, à ne pas confondre avec la précision) :

- TreeTagger (1994) : 95,7 % [Allauzen and Bonneau-Maynard, 2008]
- ME1t (2010) : près de 98 % [Denis and Sagot, 2010]



Quelle différence concrète ?

96 % d'exactitude, environ 10 mots par phrase
→ sur 10 phrases, un mot mal étiqueté dans 4 phrases

98 % d'exactitude → **deux fois moins** d'erreurs

Caractéristiques ?

Exporter et analyser

Exportez le lexique du corpus Vœux dans un tableur.

Que pouvez-vous constater concernant la répartition des fréquences de mots ?



Quelles différences avec Unitex ?

Premiers pas en CQL

Corpus Query Language

- expressions régulières : *Europe|européen.**, [] (un mot), & et | (booléens)
- neutralisations (à ajouter **après** l'expression) :
 - ▶ \c pour neutraliser la casse ("europe"\c)
 - ▶ \d pour neutraliser les diacritiques (accents, cédille)
 - ▶ etc. (voir doc)
- assistant de requête
- tri du contexte droit **et** du contexte gauche

Trier, visualiser et chercher sont 3 actions [différentes](#)

Fréquences

Index permet de chercher la fréquence d'une expression

Rechercher

Trouver en une seule recherche les fréquences de « patrie », « patriote », « patriotisme », « compatriotes »

→ permet de tester une formule de recherche (avant de se lancer en concordance)

Les vœux dans le corpus Vœux

Rechercher les vœux

Trouver en une seule recherche le souhait de « bonne année » de chaque Président

Les vœux dans le corpus Vœux

Rechercher les vœux

Trouver en une seule recherche le souhait de « bonne année » de chaque Président

`[frlemma="je"] * [frlemma="souhaiter"] * [frlemma="année"] within s`
s = dans l'espace de la phrase

`[frlemma="je"] * [frlemma="souhaiter"] * [frlemma="année"] within 25`
= dans l'espace de 25 mots

Cooccurrences

Moyen de voir comment un mot « résonne » dans un corpus

- 1 Sources
- 2 Introduction
- 3 Compter et voir en contexte
- 4 Pour finir**
 - CQFR : Ce Qu'il Faut Retenir
- 5 Bibliographie



- lexicométrie, logométrie, textométrie
- manipulations de base :
 - ▶ lexique
 - ▶ concordance
 - ▶ cooccurrences
 - ▶ index
 - ▶ CQL

Mais aussi :

- loi de Zipf
- qualité des taggers



Allauzen, A. and Bonneau-Maynard, H. (2008).

Training and evaluation of pos taggers on the french multitag corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, B. M. J. M. J. O. S. P. D. T., editor, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

<http://www.lrec-conf.org/proceedings/lrec2008/>.



Denis, P. and Sagot, B. (2010).

Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français.

In Traitement Automatique des Langues Naturelles : TALN 2010, Montréal, Canada.



Schmid, H. (1997).

New Methods in Language Processing, Studies in Computational Linguistics, chapter Probabilistic part-of-speech tagging using decision trees, pages 154–164.

UCL Press, London.