Outils logiciels pour l'analyse de corpus textuels

Bénédicte PINCEMIN
ICAR, CNRS & Université de Lyon

Ecole d'été doctorale
« Le doctorat mode(s) d'emploi »

Strasbourg,
7 juillet 2015

Objectifs

- Vue d'ensemble de ressources
  - savoir ce qui existe, + pointeurs
  - se repérer, comprendre des dépendances ou complémentarités (chaîne de traitements)
- Quelques repères méthodologiques généraux
  - notamment parce que le doctorant peut être amené à défricher un terrain que son directeur connaît peu ou pas
- Limites
  - grandes lignes, point de vue *vs* exhaustivité, dernières nouveautés

Plan

- Typologie de ressources / outils d'analyse
  - Dont : caractérisation de la textométrie
    - Dont : Positionnement de TXM et considérations sur la gratuité, l'open-source
- Constitution d'un corpus numérique
- Repères méthodologiques et éthiques

Bases : corpus + consultation

- Éléments d'appropriation
  - Possibilité de connaître et de composer le corpus
  - Outils de consultation +/- puissants, +/- orientés par un projet
- Exemples
  - Frantext : référence majoritairement littéraire pour le TLF
    - http://www.frantext.fr
  - Scientext/ScienQuest : projet d'étude linguistique d'articles scientifiques
    - http://scientext.msh-alpes.fr/scientext-site/spip.php?article1

Logiciels d'analyse

- Non liés à un corpus mais :
  - (re)connaissance disciplinaire
  - formats d'entrée et types d'analyse +/- spécifiques
  - quelquefois mise à disposition de corpus
- Exemples de portails
  - [linguistique] TGIR HumaNum / Consortium Corpus écrits / Groupe de travail Exploration de corpus
    - http://explorationdecorpus.corpusecrits.huma-num.fr/
  - [sociologie+] Méthodes qualitatives informatisées (Christophe Lejeune)
    - http://www.squash.ulg.ac.be/logiciels/index

Extraction d'informations (1)

- Application TAL (traitement automatique des langues ; NLP, natural language processing)
- Focalisation
  - Entités nommées : personnes, lieux, dates, quantités, etc.
  - Candidats termes pour lexiques spécialisés, pour peupler des ontologies
  - Repérage systématique d'informations structurées prédéfinies
    - ex. [économie] <entreprise1> <acquiert> <entreprise2> <date>
    - ex. [histoire] <personne> <nommée> <fonction> <lieu>

Extraction d'informations (2)

- Exemples de logiciels et ressources
  - Unitex (http://www-igm.univ-mlv.fr/~unitex/), Intex/Nooj (http://www.nooj-association.org/)
  - Noms propres : Univ. Tours (D. Maurel)
- Caractérisation par l'approche (Bommier-Pincemin 1999)
  - Par élection : on cherche et repère ce qu'on veut capter
    - Sémantique référentielle, à dominante nominale
    - A mettre en lien aussi avec le Web sémantique et ses ontologies/référentiels
  - Par érosion : on repère tout ce qui peut être écarté

Génération d'analyse à base de dictionnaires

- Calcul d'indicateurs prédéfinis
  - Éléments du calcul +/- transparents, +/- modifiables
  - Présentation soignée (rapport) : texte rédigé, illustration par graphiques / tableaux
- Exemples
  - Tropes (http://www.tropes.fr)
    - ex. styles argumentatif / narratif / énonciatif / descriptif
    - cf. Manuel de Tropes

    (http://www.tropes.fr/ManuelDeTropesV840.pdf)
  - Cordial Analyseur

(http://www.cordial.fr/Cordial_Analyseur/Presentation_Cordial_Analy

Génération d'analyse : exemple

- Analyse de Cordial sur l'Évangile (exemple donné par Étienne Brunet, *Écrits choisis*, t.II ch.11)

  Ce texte est accessible. Le vocabulaire est courant mais comporte quelques raretés. La complexité sémantique est plutôt élevée. Les expressions figées sont peu nombreuses. On relève une proportion de noms propres importante. Les phrases ont une longueur habituelle. Le nombre de phrases par paragraphe est très réduit. Si ce texte ne comporte pas de listes ou de titres et de sous-titres, vous devriez grouper certains paragraphes. Les phrases ont une structuration grammaticale simplifiée. Ce texte offre un niveau d'abstraction

Analyse qualitative, analyse de contenu (CAQDAS)

- Annotation
  - par le chercheur
  - de (parties de) documents (pas forcément textuels)
  - grille d'analyse personnelle (catégories)
- Rôle du logiciel
  - enregistrement unifié et organisation du corpus
  - Édition des annotations en prenant en compte les catégories
  - recherche et analyse sur les annotations → on travaille sur les annotations plutôt que sur les textes eux-mêmes

Enquêtes

- Prise en charge intégrée des différentes étapes, dont :
  - Édition du questionnaire
  - Diffusion
  - Enregistrement des réponses
  - Outils d'analyse des résultats
    - Cas particulier : questions ouvertes → analyse textuelle
- Exemples de logiciels
  - Modalisa, Sphinx Lexica

L'analyse « artisanale », à façon

- Intéresse ceux qui se débrouillent en informatique... ou qui peuvent collaborer avec un informaticien
- Traitements sur chaînes de caractères et tableaux : langages python (NLTK), perl...
- Bibliothèques statistiques (pas forcément sur texte) : R, MathLab, SPSS, SAS... avec interface graphique ou environnement intégré, notamment visualisation et analyse de réseaux (Gephi)

Data Mining, Text Mining

- Métaphore : "(énorme) masse/gisement de données non structurées", "forage de textes", "extraire des pépites", "déceler des nouveautés cachées"
- Exploitation du texte
  - Représentation généralement très appauvrie du corpus et du texte (notamment de leurs frontières et de la contextualisation induite) - idée que la quantité permet tj de trouver qqch
  - peu ou pas de retour au texte ?
- Exemples de logiciels
  - Côté open source : package tm dans R

Textométrie : pertinence pour les SHS

- point d'appui = le corpus, les contextes
  - observation de la fréquence et de la disposition des mots,
  - et surtout de leur contextualisation, locale et globale.
- travail sur des données attestées, "réelles"
  - on reste proche du texte
  - robustesse : le corpus comme une boule
- rôles : l'ordinateur mémorise et calcule, le chercheur interprète
- apports en complémentarité avec l'analyse

Textométrie : logiciels

- Généralistes
  - [Paris 3] Lexico 3 → Le Trameur
  - [Nice] Hyperbase, Hyperbase Web
  - [Lyon, Besançon,...] TXM
- Plus spécialisés
  - DTM : analyses statistiques
  - Alceste (propriétaire) ~ Iramuteq (open-source) : classification
- Internationalement, les "concordanciers"
  - ex. AntConc, WordSmith, Sketch Engine
  - Fonctionnalités : KWIC mais aussi word list

Intérêt pour TXM

- Intérêt contingent
  - multiplateformes, gratuit
  - multilingue (unicode, et dictionnaire facultatif)
  - étiquetage (analyse grammaticale automatique) à la volée
  - support (aide, maintenance) : une communauté d'utilisateurs, une équipe de développement accesible
  - (re)connu par votre discipline ?
- Motivations scientifiques
  - intérêt pour l'approche textométrique

Mutualisation, partage, recyclage, capitalisation

- développement communautaire → ouverture → standards
- ouverture (open-source)
  - pour les données comme pour les logiciels : non boîte noire
- licences de diffusion : responsabilise celui qui utilise, et portent généralement sur le devoir de citer, la possibilité de modifier, la viralité, la possibilité d'un usage commercial.
  - creative commons pour données textuelles (international, et pays par pays efforts de compatibilité du droit)

Gratuité : peut-être pas si simple ?

- Le payant : pas nécessairement inaccessible ou déraisonnable
  - +/- cher
  - licences déjà achetées par l'université
- Le gratuit : il y a naturellement des contre-parties
  - Cas 1 : esprit open-source
    - on peut vendre du service autour (formation, assistance, maintenance prioritaire...)
    - l'utilisateur est redevable : nombreuses contributions possibles, à commencer par la citation
  - Cas 2 : mode Google : droits cédés sur les données

L'open-source

- Se définit par l'accès au code, aux données, non par la gratuité
- Intérêts scientifiques
  - non boîte-noire : possibilité de comprendre le fonctionnement précis
  - évaluation : transparence (publication du code), modularité permettant comparaison méthodique de segments de traitement, diffusion
  - liberté de recherche : non dépendance à une personne ou à des intérêts particuliers, maintenance et pérennisation
  - finesse de traitement : modularité permettant

Production d'un corpus numérique

- Numérisation en mode texte
  - Par transformation de données source
  - Par récupération d'un équivalent existant
- Édition
  - Philologie : établir le texte, qualité de l'édition
  - Enrichissement
- Recours éventuel au TAL pour l'étiquetage morphosyntaxique

Obtenir du texte numérique

- Production
  - scanner + OCR
    - ex. de logiciels : ABBYY FineReader, Omnipage, Acrobat(Reader ?), services en ligne (online OCR))
    - voir licences / équipements / services locaux mutualisés (via le directeur de thèse qui peut renvoyer au labo, à la MSH...)
    - 1 image à la fois ou pas
  - reconnaissance vocale
    - dictée du chercheur, ex. dragon naturallySpeaking
    - direct depuis enregistrement : ex. Vocapia
  - transcription
    - avec un logiciel spécialisé qui facilite la saisie

Attention aux droits

- Pertinence dans le contexte de la thèse
  - risqué de capitaliser si licence pas claire
  - communication des données dans les présentations et publications
  - valorisation du travail de constitution du corpus *vs* investissement à perte
- À savoir
  - ce n'est pas parce que c'est en ligne que c'est libre
  - numériser peut déjà poser problème : droits d'auteur, droits d'éditeur
- Pour en savoir plus

Édition, philologie

- traitement de texte
  - MSWord
  - LibreOffice, OpenOffice : équivalents open-source, plus clairs pour l'encodage des caractères
- texte brut : éditeurs de textes avec chercher/remplacer élaborés
  - Notepad++, TextEdit, Fraise, Gedit, Emacs...
- XML
  - Oxygen
    - notamment pour TEI avec plugin
    - prix pour membre de consortium

TAL : étiquetage morphosyntaxique (1/2)

- Informations
  - Catégorie grammaticale : souvent notée POS « Part-of-Speech »
  - Lemme : entrée du dictionnaire

word	frpos	frlemma
Il	PRO:PER	il
en	PRO:PER	en
sera	VER:futu	être
ainsi	ADV	ainsi
,	PUN	,
désormais	ADV	désormais
.	SENT	.

- Logiciels
  - TreeTagger : open-source ; fonctionne par apprentissage
  - Cordial Analyseur : plus précis (éléments syntaxiques), payant

TAL : étiquetage morphosyntaxique (2/2)

- états de langues divers
  - oral / écrit
  - Ancien
  - SMS et nouveaux modes de communication...
    - ex. Consortium Corpus Ecrits, GT7 : Corpus d'écrits modernes et prise en compte de nouveaux modes de communication
- étape de tokenisation (découpage en mots)
  - intégrée ou non ; si non, s'assurer de la cohérence
- analyseurs syntaxiques
  - exploitation ensuite +/- fine

Positionnement du numérique dans la thèse

- À propos des statistiques textuelles, trois attitudes problématiques (Muller, 1986) :
  - fascination (scientificité, production) → avoir un usage critique
  - rejet (justifier → aller voir)
  - incapacité (accessibilité, formation)
- Mesurer son investissement car chronophage
  - temps comme première contrainte de la thèse
  - progressif : on peut déjà faire des choses avec représentation incomplète / non idéale.
  - possibilité de valoriser le corpus (par ex. en le

Usage méthodique

- Tracer
  - pouvoir revenir sur un résultat, reproductibilité
    - quelles données (archiver état)
    - quel calcul, avec quels paramètres
  - au moins expliciter, si possible justifier
- Interpréter
  - impact de la nature et du choix des données (biais)
  - herméneutique des résultats
    - Un résultat de calcul n'est pas nécessairement une réponse pertinente

"Les méthodes d'analyse factorielles [...] ont un assez grave inconvénient : elles fournissent toujours un résultat