

Panorama bref et pragmatique des outils de textométrie et apparentés

Si l'on n'a pas d'idées très précises, on peut se familiariser au domaine sur ses données avec Hyperbase ou Lexico 3. Ce sont deux outils au coeur de la textométrie, généralistes, universitaires, gratuits ou modiques, documentés, supportés par des communautés d'utilisateurs, que l'on peut expérimenter soi-même facilement.

On caractérise ci-après ces deux outils, et d'autres, par certains de leurs atouts particuliers qui peuvent aider à les singulariser (points forts mis en valeur dans l'implémentation, sans avoir nécessairement l'exclusivité bien sûr, d'autres logiciels peuvent avoir des fonctionnalités du même ordre). Pour une description plus complète de chaque logiciel il faut voir ailleurs (manuel, concepteur...).

Hyperbase : distances intertextuelles avec analyse arborée, multiplicité des fonctionnalités disponibles et dynamique de l'évolution du logiciel. Facilités d'étiquetage/lemmatisation (Cordial, TreeTagger). Indices stylométriques (richesse, accroissement lexical).

Lexico 3 : carte des sections, rebond (facilité de réutilisation des résultats d'un calcul comme base d'un autre calcul), variations possibles dans le partitionnement (regroupement de parties). Et logiciels connexes :

Lexico 2 : Calculs textométriques chronologiques disponibles (en particulier spécificités chronologiques).

MkAlign : Traitement de corpus parallèles et résonance textuelle.

Le Trameur : outil nouvelle génération issu de la même équipe (Paris 3).

Weblex : en ligne (utilisable à travers un simple navigateur – mais pour le moment import lourd pour le corpus). Moteur de recherche CQP souple et puissant. Tris multiples sur les concordances. Cooccurrences orientées. Prise en charge de corpus XML structurés, notamment étiquetés.

Taltac : nombreuses fonctionnalités pour l'analyse lexicale et textuelle d'un corpus, particulièrement pour la préparation d'un tableau à analyser (identification et reconstitution d'unités complexes). Exploitation de ressources pour enrichir de méta-informations les données textuelles. Procédure de fouille de données.

Astartex : accent mis sur la contextualisation tant globale (« macro » : textes, parties du corpus) que locale (phrase, paragraphe ou/et fenêtre) ; aide à la manipulation de la liste de vocabulaire et au parcours de la carte factorielle (index, coloration), projection géodésique. Association à Diatag cf. plus loin.

DTM-Vic : analyse des données, visualisations de type cartographique (factorielles, kohonen...) avec aides pour fiabilité de l'interprétation. Qualification des textes (parties). Particulièrement approprié aux questions ouvertes.

Alceste : classification descendante. Classes stables (par delà petites variations des définitions des contextes) et explicitation des classes par des mots et des extraits.

Xaira : moteur de recherche souple et puissant avec exploitation compétente d'un codage XML ; variations possibles dans le partitionnement (regroupement de parties).

Sato : annotation et enrichissement du corpus au fur et à mesure de l'analyse. Similarité de deux textes. Scripts de traitements élaborables à partir du « journal » (log).

Diatag : correction d'annotation, conviviale et contextuelle.

Logiciels industriels (interface soignée mais sensiblement plus onéreux) : **Sphinx-lexica**, **SPAD-T** : logiciels textométriques généralistes. **Tropes** : des calculs textométriques (rafales) mais une approche assez éloignée (analyse basée sur des catégories morpho-syntaxico-sémantique a priori, du genre 'verbes factifs', 'verbes déclaratifs', etc.).

Logiciels d'analyse textuelle et concordanciers au sens anglosaxon (nombreux ; calculs un peu différents pour leurs types et leurs formules) : **AntConc**, **MonoConc**, **Hyperpo**, **Wordsmith**, etc.

Pour l'avenir proche : le projet ANR **Textométrie** prépare une nouvelle plateforme textométrique ouverte -les contributions de tous ordres sont les bienvenues (développement, tests) : <http://textometrie.ens-lsh.fr/>

Je remercie Sergio Bolasco et Jean-Marc Leblanc, qui ont eu la gentillesse de me faire bénéficier de leurs lumières pour compléter et améliorer ce document,