Sept logiciels de textométrie

Bénédicte Pincemin

Sept logiciels de textométrie

Bénédicte PINCEMIN (CNRS, Univ. Lyon, Laboratoire IHRIM UMR 5317)

Juillet 2018

1. Contenu et organisation du document

L'approche d'analyse des données textuelles (ADT), appelée aussi textométrie, qui nous intéresse ici, est celle qui est exposée dans l'ouvrage de référence (Lebart et Salem, 1994). Elle articule les traitements qualitatifs (typiquement retour au texte et concordance) et quantitatifs (avec une place centrale du calcul des spécificités et de l'analyse des correspondances).

Les sept logiciels présentés répondent aux critères suivants :

- Ils implémentent les fonctionnalités centrales de cette approche textométrique ;
- Ce sont des logiciels gratuits facilement disponibles pour la recherche et l'enseignement ;
- Ils disposent d'une interface utilisateur graphique qui intègre et articule les différents calculs (par opposition par exemple à l'utilisation par ligne de commande ou script de bibliothèques logicielles ou *packages*, qui fournissent essentiellement des fonctions de calcul et de production de résultats).

Chaque logiciel fait l'objet d'une fiche descriptive synthétique, qui met l'accent sur ses choix de conception et ses spécialités, de façon à guider un utilisateur dans sa recherche d'un logiciel bien adapté à ses données et aux types de traitements attendus pour répondre à sa problématique.

Pour une version ultérieure du document, une description comparative analytique est prévue, sous forme de deux grilles : un tableau comparatif technique et un tableau comparatif fonctionnel.

Je remercie Ludovic Lebart et Céline Poudat, qui ont suivi activement la création de ce document, ainsi qu'Étienne Brunet, Laurent Vanni, Pierre Ratinaud, André Salem, Serge Fleury et Serge Heiden, pour leurs relectures expertes et constructives.

Documents de référence

Lebart L., Salem A. (1994). *Statistique textuelle*. Dunod, Paris. *Téléchargement :* http://www.dtmvic.com/doc/ST.html

Lebart L., Pincemin B., Poudat C. (à paraître). *Analyse des données textuelles*.

2. Fiches descriptives

2.1. DtmVic

Contact : Ludovic LEBART (Télécom-ParisTech)

Site web : **http://www.dtmvic.com**

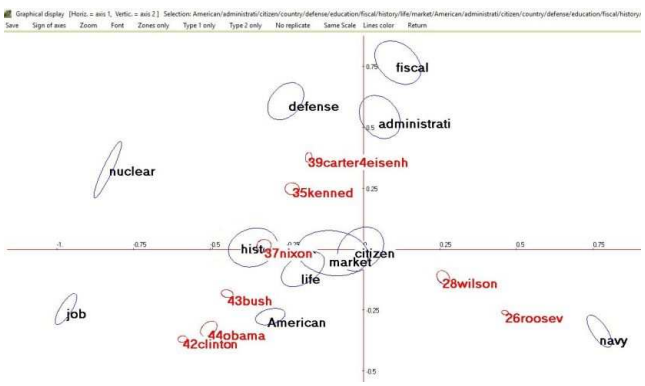

Figure 1.a : DtmVic.6.0 Analyse des Correspondances, Plan Principal simultané (lignes – colonnes)
Corpus **STATE OF THE UNION**. *Quelques présidents (en rouge) et lemmes (en noir) avec ellipses de confiance pour la précision de la position des points dans le plan (1, 2).*

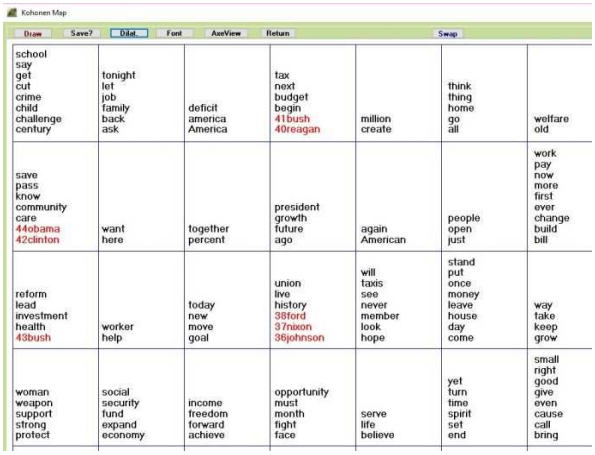

Figure 1.b : DtmVic.6.0 Carte auto-organisée simultanée (carte de Kohonen)
Corpus **STATE OF THE UNION**. *Quelques présidents (en rouge) et lemmes (en noir). Extraits d'une carte (8 × 8). Chaque case est un* cluster. *Les contiguïtés entre cases traduisent des proximités entre* clusters.

Historique et contexte de développement

DtmVic (*Data and Text Mining – Visualization, Inference, Classification*) est développé depuis les années 2000, dans la lignée du logiciel SPAD (dès les années 1980), dont le premier code, pour la partie calcul, fut publié dans l'ouvrage de Lebart, Morineau et Tabard (1977). Son contexte original de conception concerne le traitement statistique des enquêtes socio-économiques. Les possibilités de traitement statistique des questions ouvertes (en liaison avec les questions fermées) ont introduit le développement d'un volet textuel, qui s'est élargi depuis aux corpus plus généraux.

Points forts et spécialités

DtmVic offre une gamme très complète et approfondie de traitements de type statistique exploratoire multidimensionnelle (dans la lignée des travaux de Benzécri). Les techniques d'analyse par axes principaux (analyses en composantes principales, analyses des correspondances simples et multiples) (figure 1.a) et les techniques de classification (classification ascendante hiérarchique et partitionnement, classification hybride, cartes auto-organisées (SOM) de Kohonen (figure 1.b), arbre de longueur minimale, arbres additifs, sériation) sont utilisées comme des approches nécessairement complémentaires.

L'interprétation est contrôlée par des procédures de validation par rééchantillonnage avec remise (*bootstrap*) conduisant à tracer des zones de confiance (figure 1.a), qui sont systématiquement déclinées en fonction des méthodes de description (ACP, AC, ACM) et des structures de données (textes simples, enquêtes, métadonnées, etc.). Ces validations statistiques permettent d'apprécier la stabilité et donc la fiabilité des résultats observés, ce qui est en général d'autant plus important sur des textes courts, comme les réponses aux questions ouvertes, ou pour des mots/observables de faible fréquence.

DtmVic utilise les logiciels TreeTagger (Schmid, 1994) pour la lemmatisation et SplitsTree (Huson et Bryant, 2006) pour le tracé des arbres additifs.

Ressources complémentaires

Documentation multilingue (français, anglais, espagnol), jeux de données exemples disponibles, nombreux tutoriels illustrant des parcours d'analyse typiques.

Document de référence

Lebart L., Piron M. (2016). *Pratique de l'analyse des données numériques et textuelles avec Dtm-Vic,* Troisième édition, septembre 2016, version 6 de Dtm-Vic. L2C, Rivesaltes. *Téléchargement :* www.dtmvic.com/doc/DTM_Manuel_complet_2016.pdf

2.2. Hyperbase

Contact : Étienne BRUNET (Université de Nice, Laboratoire BCL UMR7320)

Site web : **http://ancilla.unice.fr**
http://logometrie.unice.fr/pages/logiciels

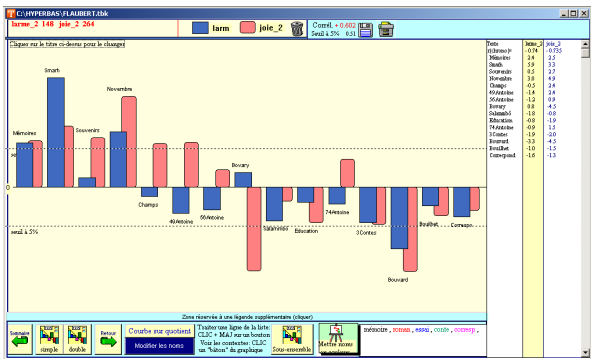

Figure 2.a : Hyperbase 9.0, Graphique
Corpus **FLAUBERT** *(version fournie avec le logiciel), visualisation de la répartition des lemmes « larme » (en bleu) et « joie » (en rouge) entre les différents textes du corpus.*

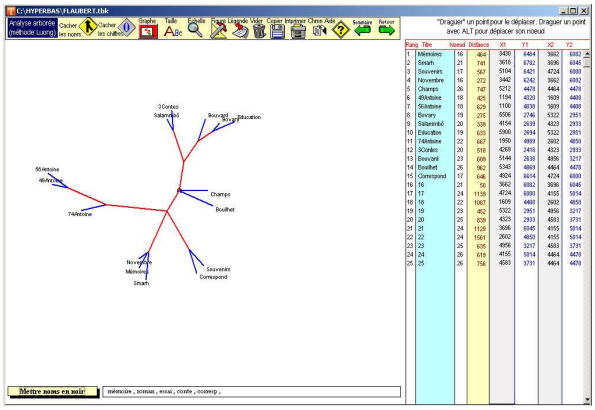

Figure 2.b : Hyperbase 9.0, Analyse arborée
Corpus **FLAUBERT** *(version fournie avec le logiciel), visualisation des proximités entre les textes du corpus (selon la distance lexicale mesurée sur les fréquences des mots)*

Historique et contexte de développement

Hyperbase a été créé en 1989 (à l'occasion d'une exposition pour le bicentenaire de la Révolution française), à partir de programmes mis au point dès le début des années 1980. D'abord utilisé pour l'analyse des textes (littéraires) du *Trésor de la langue française*, son usage au sein du laboratoire BCL s'est élargi particulièrement à l'étude de la langue et des textes latins (en collaboration avec le laboratoire LASLA à Liège) et à l'analyse linguistique des discours politiques.

Points forts et spécialités

Hyperbase se caractérise comme un logiciel particulièrement complet, intégrant la large palette de calculs que son concepteur curieux et dynamique a voulu expérimenter et offrir. Une de ses principales spécialités est le calcul et la visualisation de distances intertextuelles, avec l'analyse arborée, selon un algorithme mis au point au laboratoire (avec Xuan Luong) (figure 2.b). On notera aussi la disponibilité d'indicateurs stylométriques (comme la richesse lexicale), un calcul d'évolution du vocabulaire pour des corpus diachroniques (repérage des mots dont l'usage a tendance à augmenter ou à diminuer sur l'ensemble du corpus), et le repérage de phrases-clés (passages emblématiques). Son interface accorde une large place à la navigation hypertextuelle, elle-même résolument orientée vers le retour au texte. Le manuel, rédigé par un homme de lettres, est à la fois très agréablement écrit et très riche.

Ressources complémentaires

De très nombreuses bases de textes prêtes à l'analyse sont disponibles en ligne sur le site du logiciel et de l'équipe Logométrie, notamment des corpus d'auteurs de la littérature française (une cinquantaine), des corpus de discours politiques français (principalement présidents français depuis De Gaulle), mais aussi un corpus diachronique de textes littéraires français et un corpus de variétés du français (aires géographiques).

Document de référence

Brunet É. (2011a). *Hyperbase. Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels. Manuel de référence*. Laboratoire BCL, Université de Nice, http://ancilla.unice.fr/bases/manuel.pdf.

2.3. Hyperbase Web Edition

Contact : Laurent VANNI (CNRS, Laboratoire BCL UMR7320, Nice)

Site web : **http://hyperbase.unice.fr**

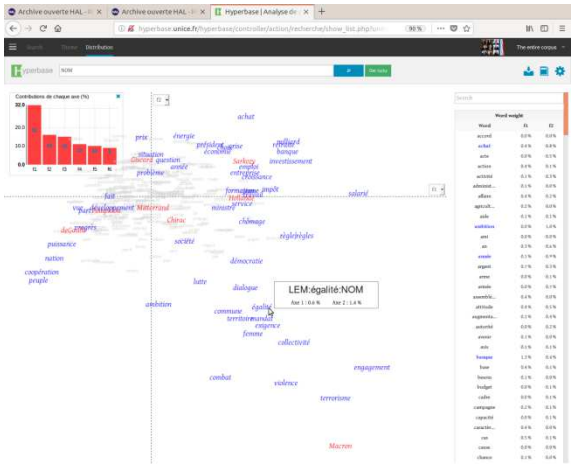

Figure 3.a : Hyperbase Web Edition (juil. 2018), Analyse des correspondances
Base **ÉLYSEE**. *Table présidents x 300 lemmes nominaux les plus fréquents (fréq. absolues), mise en évidence des mots avec contribution d'au moins 0,8 % sur l'axe 1 ou 2.*

Figure 3.b : Hyperbase Web Edition (fév. 2018), Calcul de polycooccurrence
Polycooccurrence de « lumière » dans le texte L'éducation sentimentale *du corpus* **FLAUBERT**.

Historique et contexte de développement

Le développement d'Hyperbase Web Edition a été entrepris dans les années 2010, en lien avec le concepteur de la version classique d'Hyperbase, dans l'optique d'une refonte du logiciel Hyperbase dans des technologies actuelles.

Points forts et spécialités

Hyperbase Web Edition ayant le développement le plus récent, il n'a pas encore des fonctionnalités en aussi grand nombre que les autres, mais il intègre une version des calculs les plus centraux (concordance, spécificités (figure 3.a), analyse factorielle des correspondances), une fonctionnalité-phare d'Hyperbase (l'analyse arborée), et un calcul de polycooccurrence (sur deux niveaux) doté d'une visualisation dynamique (figure 3.b). Les derniers développements s'orientent vers l'apprentissage profond (*deep learning*).

Ce logiciel se caractérise par la place centrale donnée à l'ergonomie : intuitivité de l'interface ; présentation visuelle, interactive et colorée des résultats ; facilité de mise en œuvre, car, comme son nom l'indique, Hyperbase Web Edition se présente comme un serveur en ligne, utilisable via un simple navigateur web, sans qu'il y ait d'installation logicielle à faire.

Ressources complémentaires

De nombreuses bases textuelles sont consultables, prêtes à l'emploi (héritées notamment d'Hyperbase).

2.4. IRaMuTeQ

Contact : Pierre RATINAUD (Université de Toulouse, Laboratoire LERASS EA827)

Site web : **http://www.iramuteq.org**

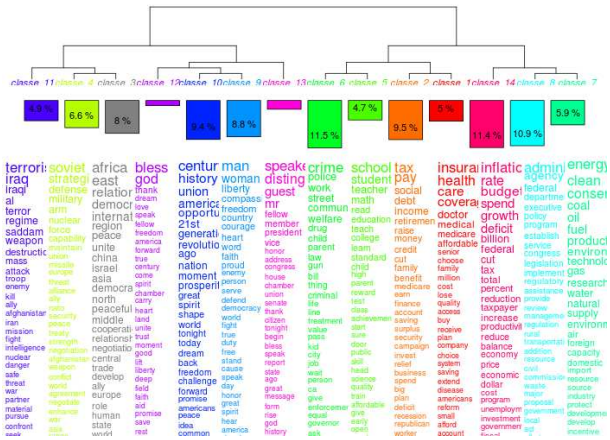

Figure 4.a : IRaMuTeQ 0.7 alpha 2, classification Reinert

Corpus **STATE OF THE UNION**, *unités de contexte de 600 caractères. Mise en évidence de 14 classes thématiques.*

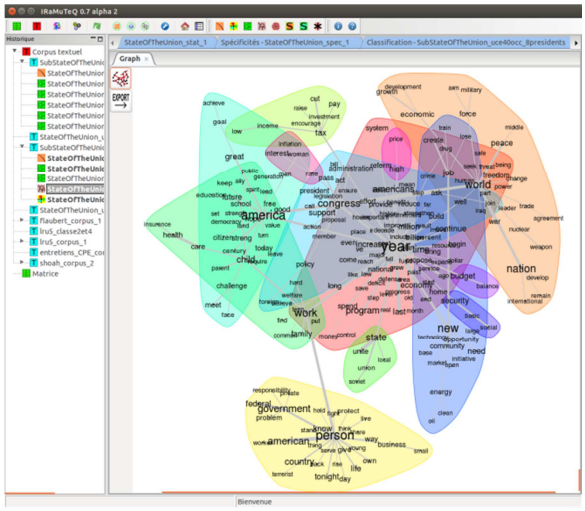

Figure 4.b : IRaMuTeQ 0.7 alpha 2, Analyse des similitudes

Corpus **STATE OF THE UNION**. *Principaux paramètres : 240 lemmes (fréq.≥ 100), Indice = cooccurrence, Présentation = graphopt, Communautés par edge.betweeness.community avec halo.*

Historique et contexte de développement

Comme l'origine de son nom l'indique, IRaMuTeQ se veut une *Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*. Il a été créé en 2010, pour implémenter en open-source la méthodologie conçue par Max Reinert et concrétisée dans le logiciel ALCESTE (diffusé par la société Image), depuis la manière de lemmatiser les mots jusqu'à l'analyse thématique de corpus. Au fil des années, le logiciel s'est cependant enrichi de fonctionnalités et visualisations nouvelles qui lui sont propres. Techniquement, il s'appuie sur l'environnement statistique **R** et sur le langage **python**.

Points forts et spécialités

La fonctionnalité d'analyse centrale est l'analyse thématique de corpus, par classification descendante de segments de textes (« classification méthode Reinert »). Pour un corpus donné, cette analyse produit automatiquement un ensemble structuré de classes (listes) de mots (figure 4.a), et de segments de textes représentatifs de chaque classe. Ces classes thématiques sont ensuite réutilisables pour colorer l'analyse des correspondances et en guider l'interprétation. IRaMuTeQ est également connu pour ses graphes arborescents de vocabulaire, par analyse des similitudes (figure 4.b). D'une façon générale sont particulièrement travaillées la qualité, la richesse informationnelle et la diversité des visualisations, avec un usage caractéristique de nombreuses couleurs, associées aux différentes classes thématiques. À noter également, pour élargir la gamme des formats d'import, IRaMuTeQ s'articule avec TXM : tout corpus qui a été importé dans l'un des logiciels est importable dans l'autre.

Ressources complémentaires

La communauté des utilisateurs est notamment animée par une liste utilisateurs dynamique (iramuteq-users@lists.sourceforge.net), permettant l'échange d'informations et l'entraide.

Document de référence

Loubère L., Ratinaud P. (2014). *Documentation IRaMuTeQ 0.6 alpha 3 version 0.1, 19 février 2014.* Toulouse, http://www.iramuteq.org/documentation/fichiers/documentation_19_02_2014.pdf

Ratinaud P., Déjean S. (2009). IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. In *Colloque Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009)*, Toulouse, http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation_mashs2009.pdf/view

2.5. Lexico 5

Contact : André SALEM (Université Paris 3)

Site web : **http://www.lexi-co.com**

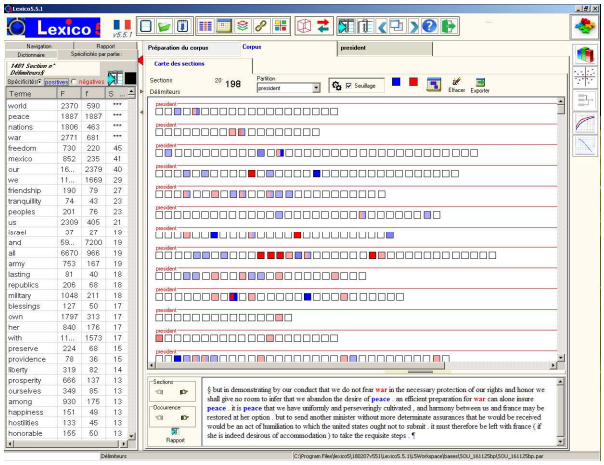

Figure 5.a : Lexico 5.5.1, Carte des sections
Corpus **STATE OF THE UNION**. *Mot « peace » en bleu, « war » en rouge. À gauche, cooccurrents de « peace ».*

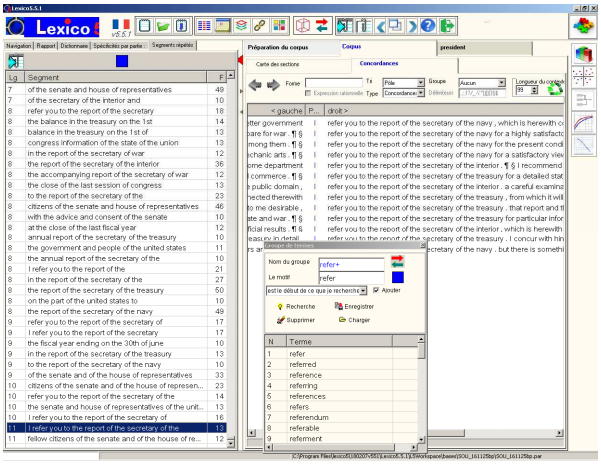

Figure 5.b : Lexico 5.5.1, Segments Répétés et Tgen
Corpus **STATE OF THE UNION**. *Concordance d'un long segment répété et construction d'un groupe de termes commençant par « refer »*

Historique et contexte de développement

Lexico 5 est développé depuis le milieu des années 2010. Il fait suite à Lexico 3 développé à partir de 2003. Lexico 3 était lui-même précédé de Lexico 1 puis Lexico 2, développements débutés dans les années 1980 dans le contexte du laboratoire de Saint-Cloud – acteur majeur de la création, de la théorisation et de la diffusion de la lexicométrie – et diffusés surtout à partir des années 1990.

Points forts et spécialités

Plusieurs fonctionnalités originales ou plus développées correspondent aux innovations méthodologiques du concepteur : interface de définition d'objet de recherche dite TGEN (type généralisé) permettant la construction et la réutilisation souples d'une liste de mots (figure

5.b) (Lamalle & Salem 2002) ; calcul des segments répétés (figure 5.b), intégration simple de ces segments comme unités pour les analyses ultérieures, et inventaires distributionnels (Salem, 1984, 1987) ; et surtout peut-être (en termes d'originalité dans l'état de l'art des logiciels actuels), spécificités chronologiques pour l'analyse de séries temporelles (Salem, 1988, 1991). En outre, la visualisation par « carte des sections » (figure 5.a) est une fonctionnalité centrale, qui permet d'articuler une vue synthétique globale de la répartition d'un mot (ou groupe de mots) au fil du corpus, et l'étude des cooccurrences.

Travaillant sur du texte brut (sans lemmatisation ni étiquetage morphosyntaxique – ou alors en traitant séparément chacune de ces informations : textes réécrits comme une suite de lemmes, ou comme une suite d'étiquettes), le logiciel est capable de traiter de façon efficace et rapide de gros corpus (au moins plusieurs dizaines de millions d'occurrences). Une innovation de Lexico 5 par rapport à Lexico 3 est la prise en charge de corpus en unicode, étendant au plus large l'éventail des langues pouvant être prises en charge.

Ressources complémentaires

Un recueil d'une vingtaine d'*explorations textométriques* (**http://www.tal.univ-paris3.fr/lexico/navigations-tdm.xml**) donne des exemples diversifiés d'études réalisées avec Lexico 3, illustrant concrètement des démarches méthodiques d'analyse de corpus.

Document de référence

Lamalle C., Martinez W., Fleury S., Salem A., Fracchiolla B., Kuncova A. & Maisondieu A. (2003). *Lexico 3 version 3.41 février 03. Outils de statistique textuelle. Manuel d'Utilisation.* Laboratoire SYLED-CLA2T, Université de la Sorbonne nouvelle - Paris 3, http://www.lexi-co.com/ressources/manuel-3.41.pdf.

2.6. Le Trameur

Contact : Serge FLEURY (Université de Paris 3, Laboratoire CLESTHIA EA7345)

Site web : **http://www.tal.univ-paris3.fr/trameur**
http://www.tal.univ-paris3.fr/trameur/iTrameur/

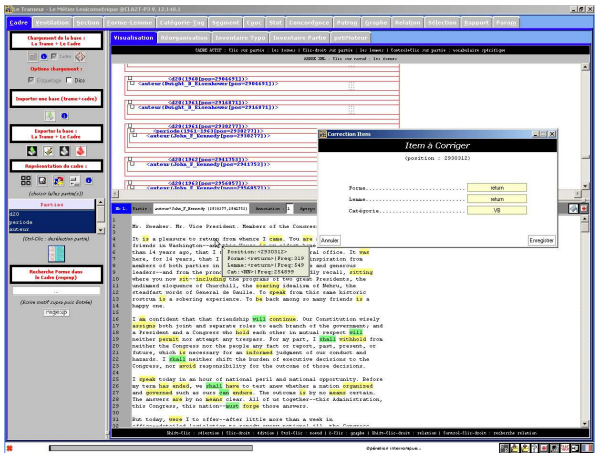

Figure 6.a : Trameur 12.148.1, Visualisation du Cadre, Coloration, Correction d'un item
Corpus **STATE OF THE UNION**. *On a choisi de surligner les verbes en jaune et les modaux en vert. On repère une occurrence du verbe*

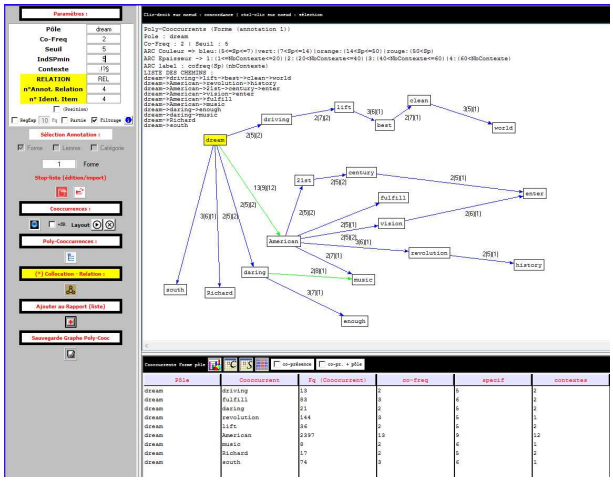

Figure 6.b : Trameur 12.148.1, Graphe de poly-cooccurrence
Corpus **STATE OF THE UNION**. *Poly-cooccurrents du mot « dream ».*

« return » qui n'a pas été surlignée (faussement étiquetée nom), on corrige son étiquette

Historique et contexte de développement

Le Trameur a été conçu sur la base du modèle proposé par André Salem dans *Pour une textométrie opérationnelle* (Söze-Duval 2008), qui propose une généralisation du modèle textométrique sous la forme d'une « trame » de mots localisés dans un « cadre » délimitant des zones de contexte (figure 6.a). Le Trameur s'inscrit dans la tradition des développements de textométrie à Paris 3, principalement Lexico 3, dont il reprend la plupart des modules : spécificités, analyse des correspondances, concordance, carte de sections, etc., et inclut aussi la polycooccurrence de l'outil CooCS (Martinez 2003) (figure 6.b). (En revanche les spécificités chronologiques restent une originalité de Lexico.)

Points forts et spécialités

Dans ce contexte de l'équipe de Paris 3, la principale nouveauté du Trameur consiste à pouvoir travailler sur des textes structurés et annotés (dont XML). En effet, un contexte majeur d'expérimentation du Trameur a été le projet Rhapsodie (ANR-07-CORP-030), autour d'un corpus annoté en arbres syntaxiques et prosodiques. Notamment le Trameur permet la modification, l'ajout et la fusion d'annotations sur les mots en cours d'analyse (figure 6.a). Dans ses choix graphiques, le Trameur propose souvent des présentations visuelles de ses résultats sous forme de graphes (figure 6.b), quitte à produire un affichage chargé si les résultats sont nombreux (la recherche doit alors être davantage ciblée). L'interface d'interrogation privilégie des formulaires (plutôt qu'un langage d'interrogation), et donne des possibilités de recours aux expressions régulières pour davantage de puissance expressive.

Le Trameur est développé par le même concepteur que mkAlign, dédié à l'analyse textométrique de corpus parallèles alignés. Ces deux logiciels ont une évolution convergente, si bien que le Trameur s'enrichit progressivement de fonctionnalités spécialisées pour les corpus alignés, comme la résonance textuelle (Salem 2004).

Depuis décembre 2016 est développée une version en ligne du Trameur, appelée iTrameur, avec de nouveaux enrichissements des fonctionnalités.

Ressources complémentaires

La page web de l'outil est particulièrement fournie et rassemble des ressources en tous genres, des jeux de données aux liens vers les textes intégraux des références bibliographiques. Le manuel est tenu constamment à jour.

Document de référence

Fleury S. (2017). *Le Trameur. Manuel d'utilisation.* CLA2T / SYLED, Université Sorbonne nouvelle Paris 3, http://www.tal.univ-paris3.fr/trameur/leMetierLexicometrique.pdf

Fleury S., Zimina M. (2014). Trameur: A Framework for Annotated Text Corpora Exploration. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, August 2014, Dublin City University and Association for Computational Linguistics, Dublin, Ireland : 57-61.

Söze-Duval K. (2008). *Pour une textométrie opérationnelle*. Inédit.
http://www.tal.univ-paris3.fr/trameur/RTI6provisoire.doc

2.7. TXM

Contact : Serge HEIDEN (ENS de Lyon, Laboratoire IHRIM UMR5317)

Site web : **http://textometrie.org**
http://textometrie.ens-lyon.fr

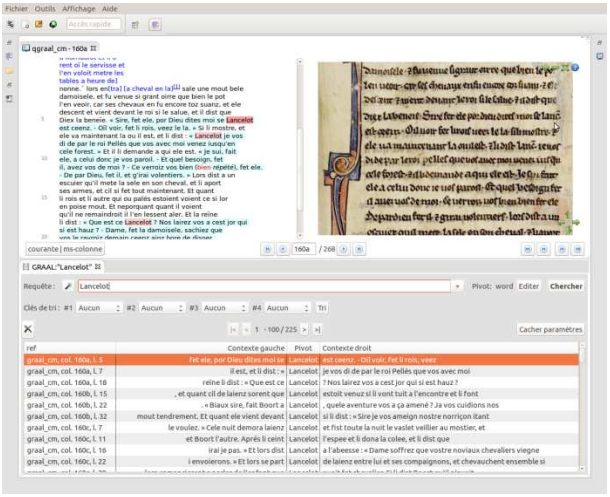

Figure 7.a : TXM 0.7.9, Concordance avec retour au texte et vue synoptique du document source

Corpus **GRAAL**. *Une concordance est calculée sur « Lancelot ». Un double-clic sur une ligne de la concordance permet de voir l'occurrence dans le contexte de la page, et le passage du manuscrit correspondant.*

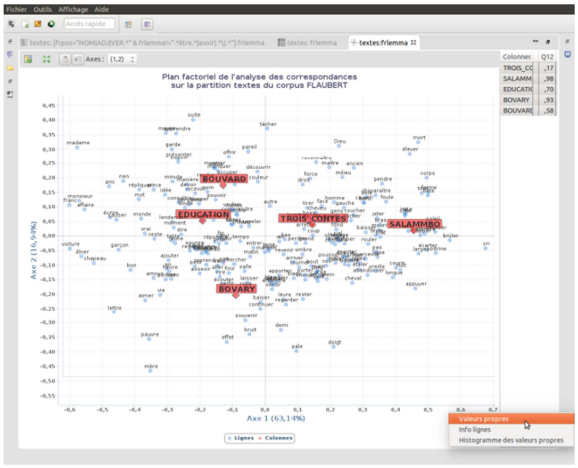

Figure 7.b : TXM 0.7.8, AC

Corpus **FLAUBERT**. *Un index puis une table lexicale ont permis de sélectionner 290 lemmes de noms, verbes ou adjectifs parmi les 300 plus fréquents (en excluant auxiliaires, erreurs d'étiquetage, mots quasi-exclusifs). La macro CAFilter a permis d'alléger le graphique des points à faible cos² et faible contribution dans le plan. Zoom et sélection de points clarifient l'exploration.*

Historique et contexte de développement

TXM a été initié dans le cadre du projet ANR Textométrie (ANR-06-CORP-029), réunissant les équipes de Lyon, Nice, Paris et Besançon. Le développement a débuté en 2009, avec le double objectif de repenser les calculs textométriques dans le contexte des corpus annotés et structurés, et de concevoir un logiciel modulaire open-source. Ainsi le développement de TXM adopte les standards internationaux du domaine (Unicode, XML, TEI pour les sources textuelles et Java, OSGi, XSLT2 pour les composants logiciels) et s'appuie sur des composants open-source de référence (moteur de recherche CQP, environnement statistique **R**, dont les librairies **factomineR** (Husson *et al.*, 2009) et **textometry**, plateforme Eclipse RCP, plateforme web GWT). Le projet se poursuit depuis 2013 sous la forme d'un co-développement impliquant Lyon (IHRIM UMR5317) et Besançon (ELLIADD EA4661).

Points forts et spécialités

Une caractéristique générale du logiciel est la recherche de souplesse et de puissance. Un effort important a été consenti pour la prise en charge de formats de corpus très divers, du plus simple (copier/coller, txt) au plus riche (XML-TEI). TXM offre ensuite une exploitation très complète des informations disponibles dans le corpus (définition dynamique des références de localisation et tris à de multiples niveaux dans les concordances, construction

des tableaux de données et ajustement de leurs marges pour les calculs statistiques (figure 7.b), etc.), avec le langage d'interrogation CQL qui permet l'expression de requêtes très précises. Par ailleurs, une attention particulière est portée à la restitution des documents source (*philologie numérique*, cf. Guillot *et al*., 2016) (figure 7.a), avec une ouverture vers le multimédia. L'interactivité des sorties graphiques est aussi à souligner (zoom dans les analyses des correspondances (figure 7.b), retour au texte depuis un point d'une courbe de progression, etc.), et les dernières évolutions introduisent l'annotation en cours d'analyse. Le logiciel peut être piloté et personnalisé par des scripts, et les résultats exportés dans des formats standards exploitables par d'autres outils. Il se décline en deux versions, pour ordinateur personnel et pour serveur Web.

Ressources complémentaires

La communauté utilisateurs est animée par une liste de diffusion (**txm-users** ; liste anglophone : **txm-open**) et un wiki (**https://groupes.renater.fr/wiki/txm-users**).

Document de référence

Heiden S., Decorde M., Jacquot S. & Pincemin B. (2018). *Manuel de TXM, Version 0.7.9 Février 2018*. ENS de Lyon & Université de Franche-Comté, http://textometrie.ens-lyon.fr/files/documentation/Manuel%20de%20TXM%200.7%20FR.pdf

Heiden, S., Magué, J-P. & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Bolasco S., Chiari I., Giuliano L. editors, *Statistical Analysis of Textual Data. Proc. of JADT 2010 (10th International Conference on the Statistical Analysis of Textual Data)*, Edizioni Universitarie di Lettere Economia Diritto, Roma : 1021-1032.