Obvie: interface web pour la fouille et la comparaison de textes

Motasem Alrahabi

Obvie: interface web pour la fouille et la comparaison de textes

Motasem Alrahabi

ObTIC-SCAI, Sorbonne Université, 75005 Paris
motasem.alrahabi@gmail.com

Résumé. Obvie est un moteur de recherche qui permet d'explorer les corpus textuels avec des fonctionnalités avancées de fouille, de statistiques lexicales et de comparaison. Nous présentons dans cet article les principales fonctionnalités de cette application avec un scénario de fouille sur un corpus textuel issu du domaine socio-médical.

1. Présentation

Obvie[1] a été développé dans le cadre du Labex OBVIL[2] et de sa suite, l'équipe-projet ObTIC[3], pour répondre à des besoins spécifiques en termes d'exploration et de comparaison de textes qui ne se limitent pas au domaine littéraire. Notre conception de l'application devait répondre aux exigences suivantes : valoriser les corpus numérisés et structurés en XML-TEI et édités au sein de la bibliothèque numérique de l'équipe[4] ; prendre en compte la diversité des domaines de spécialités des corpus traités ; mettre à disposition de la communauté des humanités numériques une plateforme ouverte et générique accueillant les corpus existants, mais permettant également de traiter de nouveaux textes.

L'architecture générale de l'application est basée sur une suite de fonctionnalités qui offrent des dispositifs de lecture et d'interprétation avancées.

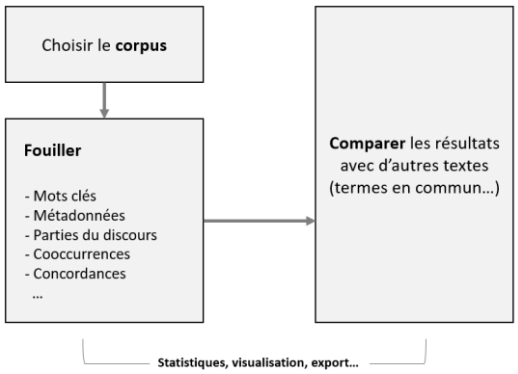

FIG. 1 - *Architecture conceptuelle de la plateforme Obvie*

[1] https://obvil.huma-num.fr/obvie
[2] https://obvil.sorbonne-universite.fr
[3] https://obtic.sorbonne-universite.fr
[4] https://obvil.sorbonne-universite.fr/bibliotheque

À partir d'une première requête et, éventuellement, d'un filtrage par métadonnées (auteur, date et titre), les résultats sont présentés selon une série de « vues » :

- La vue « Corpus », dans laquelle sont listés tous les fichiers du corpus qui répondent à la requête. Il est possible de sélectionner une partie des résultats afin de créer un sous-corpus pour la suite.
- Les vues « Fréquences » et « Nuage », qui montrent les mots cooccurrents des termes de la requête, présentés (respectivement) dans un tableau dont les résultats peuvent être triés ou sous forme de nuage de mots. Les résultats sont filtrés par nombre ou par catégorie grammaticale[5]. L'utilisateur a le moyen également de télécharger ces données au format CSV.
- Les vues « Extraits », « Concordance » et « Document » : afficher les passages qui répondent à la requête, dans un contexte large, sous forme de concordance ou dans le contexte du document d'origine.
- La vue « Comparaison ». La plateforme offre le moyen de comparer un texte répondant à la requête avec les autres textes du même corpus. Cette fonctionnalité, très utile pour des analyses de similarité, de reprise ou d'emprunt, permet de visualiser les deux documents côte à côte et de surligner les mots fréquents de chaque document, les noms propres[6] cités dans chaque document et les noms en commun entre les deux documents.

FIG. 2 – Le menu des vues dans Obvie

Grâce à toutes ces fonctionnalités, l'utilisateur a le moyen d'explorer rapidement le contenu de son corpus[7], et d'en extraire des informations pertinentes. Obvie est utilisé dans différents travaux d'enseignement et de recherche sur des textes littéraires[8]. Cependant, l'outil fonctionne sur d'autres genres textuels. À ce titre, nous présentons un scénario de fouille d'un corpus issu du domaine médico-social.

[5] L'application est dotée d'un moteur de lemmatisation (https://obvil.huma-num.fr/alix/) qui permet de rechercher par lemmes et de filtrer les résultats par catégorie grammaticale (adjectif, adverbe, etc.).

[6] Les noms propres sont généralement obtenus grâce à des dictionnaires construits manuellement et à des expressions régulières.

[7] Pour indexer un nouveau corpus, il suffit de nous envoyer un corpus en TEI qui suit le schéma Teinte (https://github.com/oeuvres/Teinte). Une mise à jour est prévue pour permettre aux utilisateurs d'indexer eux-mêmes leurs corpus (web service).

[8] Obvie est utilisé en complémentarité avec d'autres outils de fouille développés par l'équipe ou par des partenaires, comme Ariane (Alrahabi 2021) et Philologic (Allen *et al.* 2013).

2. Scénario de fouille

Nous présentons dans cette section un parcours de fouille textuelle exploitant les fonctionnalités de la plateforme Obvie. Il s'agit d'une analyse d'un corpus issu du domaine socio-médical, avec l'objectif d'appréhender le contenu de celui-ci par une approche à la fois quantitative que qualitative.

2.1 Présentation des données

Le corpus du Samu Social est le fruit d'une initiative de la faculté de Médecine de Sorbonne Université, qui a mis en place des stages d'immersion dans le monde professionnel (urgences, SAMU Social, Sida info service, etc.) à destination de ses étudiants. Dans les stages "courts" du Samu Social (2010-2020), les étudiants devaient accompagner les équipes dans une maraude nocturne (double écoute et rencontre des sans-abris) et répondre à l'issu de cette expérience à une enquête avec des questions ouvertes.

Dans le cadre de ce projet de recherche, le travail sur ces textes a pour objectifs d'analyser le vécu socio-médical pour une meilleure prise charge médicale et d'évaluer l'expérience vécue dans la formation universitaire (Alexandre *et al*. 2022). Le corpus est composé de 2299 récits anonymisés (environ 150 mots par récit), et chaque texte est accompagné de plusieurs métadonnées : âge, genre, milieu social, expérience humanitaire, date, etc.

2.2 Analyse et résultats

Avec Obvie, nous avons entrepris une suite d'analyses, quantitative et qualitative, autour des termes les plus significatifs du corpus. Une première exploration des textes par catégories grammaticales nous permet d'appréhender les thématiques majeures du corpus. Parmi les 20 premiers substantifs que la vue Fréquences propose, nous trouvons les termes représentatifs suivants : nuit (2263), garde (1801), équipe (1639), expérience (1463), maraude (1243), rue (1134), centre (838), hébergement (625), abri (462), urgence (462), etc.

Filtrer dans la même vue par les noms propres nous place le terme SDF en tête de liste avec 553 occurrences. Nous relevons à ce stade le faible nombre d'observations médicales, de diagnostics et d'avis médicaux présents dans les rapports (le terme « diagnostic », par exemple, apparait 5 fois seulement dans les 2299 textes). Cela tient sans doute aux conditions du stage,s qui réduit l'externe au rôle d'observateur.

Quant aux verbes, les plus fréquents servent de marqueurs d'observation, de perception et de réflexion (*voir*, *écouter*, *trouver*, *découvrir*, *penser*, *comprendre*, etc.). Ces marqueurs reflètent un aspect important de l'expérience qui est vue comme un lieu d'apprentissage et de découverte. Une lecture attentive des résultats dans la vue Extraits confirme que ces verbes portent souvent sur des substantifs déjà relevés[9] : la rue, les urgences, les équipes, etc. Exemple :

[9] Soulignons que notre approche lexicale ici ne permet pas d'analyser automatiquement le contexte. D'où le besoin d'une lecture attentive des résultats : les fréquences affichées des adjectifs doivent être minutieusement vérifiées afin de bien identifier l'objet des jugements

J'ai pu comprendre le fonctionnement interne et les missions du samu social grâce à une équipe très disponible qui répondait à toutes mes questions (#1041).

En ce qui concerne les adjectifs, nous avons sélectionné à partir des résultats les éléments les plus fréquents et qui sont positivement ou négativement connotés :

Positif	Fréquence	Négatif	Fréquence
enrichissant	742	difficile	309
intéressant	669	démuni	136
bon	661	dur	126
humain	580	défavorisé	85
sympathique	197	précaire	79
accueillant	190	frustrant	66
beau	172	compliqué	55
utile	163	mauvais	47
sympa	125	triste	39
agréable	124	agressif	33

TAB 1 – *Les adjectifs connotés les plus fréquents (valeurs absolues)*

Comme nous pouvons le constater, les adjectifs en apparence négatifs sont moins nombreux que les adjectifs positifs, ce qui reflète généralement le caractère plutôt positif de l'expérience vécue par les jeunes externes en médecine. Cliquer sur un résultat (*intéressant* par exemple) nous amène au concordancier, puis au contexte d'origine du terme dans le document :

J'ai trouvé cela très intéressant de nous ouvrir aux métiers parallèles à une activité médicale (#1026).

Il est également possible d'explorer l'apparition de ces adjectifs tout au long de la période couverte par le corpus, grâce à la frise chronologique interactive. Par exemple, pour une requête combinant deux des adjectifs opposés les plus fréquents (*intéressant* et *difficile*), la frise nous permet d'observer que *difficile* dépasse *intéressant* à un seul moment de la période, soit en 2017 :

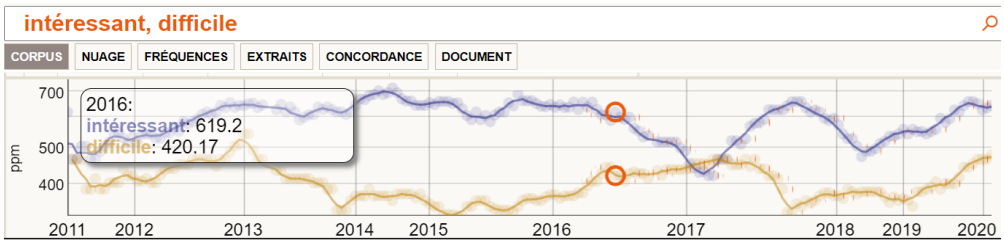

FIG. 3 – La frise chronologique montrant les fréquences de deux termes recherchés

(la garde, la double écoute, la maraude, la relation à l'équipe, la rencontre des sans-abris, etc.) et les sources ou les émetteurs de ces jugements (l'étudiant, les autres membres de l'équipe, les SDF, etc.).

Signalons enfin l'intérêt de filtrer les fréquences selon la catégorie grammaticale des adverbes. En dehors de quelques adverbes évaluatifs (*malheureusement, agréablement*…), nous constatons que la majorité des adverbes est de nature intensive (*particulièrement, notamment, forcément, réellement, énormément,* etc.). Les étudiants ont recours à ce type de marqueurs pour renforcer leurs opinions et observations :

Les équipes sont particulièrement à l'écoute et pleine de gentillesse (#1019)

Nous terminons l'exploration du corpus par la comparaison de textes. Examinons le premier résultat retourné par le système pour la requête SDF. Obvie permet d'identifier les passages dans lesquels ce mot apparaît, puis, pour chacun d'eux, d'afficher les autres passages du corpus qui présentent la plus grande similarité lexicale, calculée selon un ordre de pertinence[10]. En choisissant ensuite d'afficher la comparaison texte à texte, il devient possible d'identifier les termes communs aux deux passages et leur répartition.

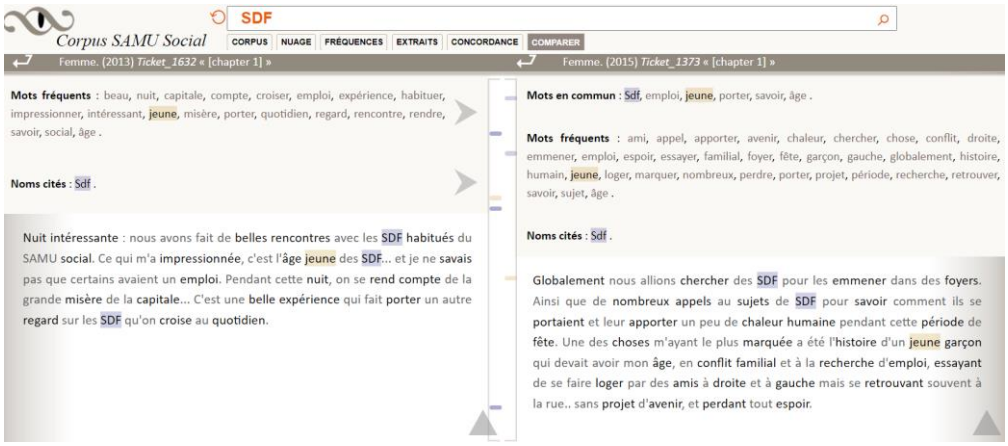

FIG. 4 – Interface d'Obvie : *la vue « Comparaison » entre deux textes*

Dans cet exemple, il apparaît clairement que la jeunesse des SDF est une expérience commune aux deux externes dont les textes ont été rapprochés.

3. Conclusion

Si Obvie a initialement été conçue pour l'exploration de textes littéraires, l'ensemble de fonctionnalités qu'elle offre, notamment la recherche par catégories grammaticales, les cooccurrences et la comparaison, en font un moyen rapide d'exploration de corpus couvrant des domaines très différents, y compris dans le domaine des humanités médicales. Elle permet de confirmer ou d'infirmer des hypothèses de recherche, mais peut aussi mettre en lumière des aspects d'un corpus que l'analyse traditionnelle ne révèlerait pas forcément tout de suite. Cela fait d'Obvie un outil d'analyse de corpus performant.

[10] Le principe consiste à pondérer le nombre de d'occurrences trouvées dans des textes, la taille de ces textes, et le nombre de documents trouvés (l'algorithme Okapi BM25).

Références

Alexandre D., Duguet A., Alrahabi M., Renaud M.-C., Riguet M., Gay F. (2022), Le médical et le social. Analyse sémantique des rapports de l'immersion d'étudiants de médecine dans le Samu social, in *Humanités Numériques Littéraires*, sous la dir. de Didier Alexandre, Paris, Classiques Garnier, 2022 (*à paraître*)

Allen T., Gladstone C., Whaling R. (2013), PhiloLogic4: an Abstract TEI Query System, Journal of the Text Encoding Initiative, Issue 5, June 2013.

Alrahabi M., Ariane: dispositif de fouille et de lecture synthétique de textes, Actes de DigitAl Humanities and cuLtural herItAge: data and knowledge management and analysis (Atelier Dahlia), Jan 2021, Montpellier, France

Summary

Obvie is a search engine that makes it possible to explore textual corpora with advanced search features, lexical statistics and textual comparison. In this article, we present the main features of this application with a typical search scenario on a corpus from the domain of social medicine.