# *Analyse contrastive de motifs issus des Cahiers Citoyens*

**Mots clés :** débat public, sémantique computationnelle, modèle de langue, plongement, textométrie

## 1 Contexte

À la suite de l'émergence du mouvement des Gilets jaunes fin 2018, des Cahiers de doléances, renommés ensuite Cahiers citoyens (CC), ont été déposés dans les mairies et remplis librement par les personnes qui le souhaitaient. Quantitativement, les plus de 19 000 Cahiers (environ 40 millions de mots comptés à partir des transcriptions) constituent la composante la plus importante du corpus du Grand débat national (GDN) janvier-mars 2019.

Si le corpus contenant les contributions déposées sur des plates-formes du GDN a d'ores et déjà fait l'objet de divers travaux en traitement automatique des langues : repérage des thèmes abordés, extraction de réseaux lexicaux, repérage de motifs ou phrases-type…, celui des Cahiers citoyens a principalement été étudié par des approches en SHS qui restent peu automatisées et relatives à des aires géographiques circonscrites [1]. L'analyse systématique du corpus CC permettrait de contraster les thématiques abordées dans les contributions issues des plates-formes du GDN et celles du CC et ainsi de mesurer l'impact du dispositif de consultation sur les profils des contributions.

## 2 Sujet

L'objectif du stage est de fournir une analyse des contributions du corpus CC en termes de thèmes abordés. Cette analyse sera fondée sur l'analyse sémantique computationnelle des contributions [Ji et al., 2008]. Différentes approches seront utilisées : extraction de réseaux lexicaux et le repérage de motifs ou phrases-type, classification par un modèle géométrique [Ploux et al., 2021] ou un modèle de langue comme CamemBERT [Martin et al., 2019] et des plongements lexicaux [Park, 2018] fournis par ce modèle. Les analyses seront développées pour s'appliquer à l'ensemble du corpus ou à des sous-corpus définis par des critères

1. Une synthèse a été réalisée par la société Cognito, elle est disponible sur le site du GDN.

textométriques, thématiques, géographiques et/ou socio-démographiques.

Le stage comportera les étapes suivantes (une attention particulière sera portée aux conditions de réutilisabilité des ressources et codes produits, et donc à leur documentation tout au long du stage) :

- appropriation des travaux déjà réalisés sur l'analyse des corpus GDN et CC ;
- analyse textométrique du corpus CC à partir des thèmes déjà retenus (et des termes associés) pour l'analyse du GDN ;
- rédaction d'un état de l'art concernant les modèles de langue, le réentraînement de ces modèles, les modèles fondés sur les graphes, l'utilisation des plongements lexicaux et les algorithmes développés pour la comparaison et la classification de ces vecteurs ;
- mise en place d'analyses à partir de modèles existants
- étude contrastive des résultats obtenus à partir du corpus ou de sous-corpus des CC et de celui de la plate-forme du GDN ;
- rédaction du rapport de stage, et mise en forme des ressources et codes produits.

## Références


[Ji et al., 2008] Ji, H., Lemaire, B., Choo, H., and Ploux, S. (2008). Testing the cognitive relevance of a geometric model on a word association task : A comparison of humans, acom, and lsa. *Behavior research methods*, 40(4) :926–934.

[Martin et al., 2019] Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de La Clergerie, É. V., Seddah, D., and Sagot, B. (2019). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*.

[Park, 2018] Park, J. (2018). L'optimisation du plongement de mots pour le français : une application de la classification des phrases (optimization of word embeddings for French : an application of sentence classification). In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pages 281–292, Rennes, France. ATALA.

[Ploux et al., 2021] Ploux, S., Genay, M., and Ploux-Chillès, L. (2021). Les mots du grand débat national : les réseaux lexicaux des contributions déposées sur trois plateformes. *Humanités numériques*, (4).


## 3 Formation requise

Ce stage s'adresse aux étudiant.e.s de master 2 en informatique/analyse de données ou en TAL avec une formation suffisante pour l'utilisation autonome d'un langage de programmation (de préférence Python et R) et d'outils de TAL (outils fondés sur l'apprentissage, modèles de langue, classifieurs, si possible outils statistiques de lexicométrie).

## 4 Lieu du stage

Le stage se déroulera principalement au Laboratoire en sciences et technologies de l'information géographique (LaSTIG) sur le site de l'Institut national de l'information géographique et forestière (IGN) à Saint-Mandé ; des réunions de travail seront organisées régulièrement au laboratoire CAMS à l'EHESS à Paris (6ème).
Laboratoire en sciences et technologies de l'information géographique
Institut national de l'information géographique et forestière
73 avenue de Paris
94165 Saint-Mandé Cedex
métro : Saint-Mandé - ligne 1 ou RER A - Vincennes

Centre d'analyse et de mathématique sociales
École des hautes études en sciences sociales
54 boulevard Raspail
75006 Paris

## 5 Durée et rémunération

durée : entre 5 et 6 mois
début possible à partir d'avril 2023
gratification au taux horaire net de 4,05 €

## 6 Encadrement du stage

Catherine Dominguès, chercheure HDR au LaSTIG en TAL et géomatique,
catherine.domingues@ign.fr
Sabine Ploux, chercheur HDR au CAMS EHESS-CNRS, en linguistique computationnelle,
sabine.ploux@ehess.fr

## 7 Pour candidater

Des entretiens seront organisés. Préalablement, un dossier de candidature est à envoyer aux encadrantes et devra contenir les documents suivants : CV, lettre de motivation, derniers relevés de notes (M1, et premier semestre de M2 si possible), description des enseignements suivis (un lien vers le site internet de la formation est le bienvenu), dernier mémoire ou rapport de stage.