1. État de l'art : corpus politiques français et analyse du lexique idéologique

1.1. Tradition française : langages du politique et lexicométrie

En France, l'étude des "langages du politique" s'est structurée dès les années 1980 autour de deux traditions très articulées :

- **l'analyse du discours politique**, qui met l'accent sur les formations discursives, les positionnements idéologiques, les genres (meeting, débat, tract, tweet…) ;
- **la lexicométrie / textométrie**, qui outille cette analyse par des corpus, des comptages et des visualisations. (Persee)

La revue *Mots. Les langages du politique* joue un rôle central dans cette histoire : elle documente à la fois la construction de corpus politiques (discours parlementaires, professions de foi, tracts, tweets…) et les méthodes d'analyse associées (statistiques textuelles, cooccurrences, typologies thématiques, etc.). (Persee)

Sur le plan strictement lexical, les travaux de **Dubois** et du **Centre de lexicologie politique** sur *Le vocabulaire politique et social en France de 1869 à 1872* constituent un jalon fondateur : ils posent déjà la question de ce que signifie "vocabulaire politique" et de la manière de circonscrire un champ lexical politique pertinent pour l'analyse historique et sociale. (OpenEdition Books)

Plus récemment, **Sylvianne Rémi-Giraud** propose une mise au point théorique sur les liens entre **sémantique lexicale** et **langages du politique**, en soulignant le paradoxe d'une relation à la fois nécessaire et méthodologiquement difficile : comment traiter les champs lexicaux, la polysémie, les glissements de sens, sans écraser la dimension discursive et idéologique ? (OpenEdition Journals)

Enfin, des auteurs comme **Pincemin** insistent sur l'**hétérogénéité des corpus** (périodes, genres, supports, acteurs) et sur les effets que cette hétérogénéité produit sur les traitements lexicométriques – question directement pertinente pour un projet multi-corpus comme le tien. (perso.liris.cnrs.fr)

1.2. Corpus politiques français existants

1.2.1. Corpus parlementaires et discours institutionnels

Plusieurs projets ont constitué des **corpus de débats parlementaires** français encodés selon des standards (SGML puis TEI), avec un niveau de structuration élevé (sessions, tours de parole, interruptions, métadonnées sur les orateurs, etc.) :

- les travaux de **Serge Heiden** sur l'**encodage uniforme et normalisé** de débats parlementaires, qui servent de base à des corpus TEI exploitables en lexicométrie ; (Persee)
- le corpus **TAPS-fr** (Transcription and Annotation of Parliamentary Speech), présenté dans le cadre de CLARIN / LREC, qui fournit un corpus monitor de débats de l'Assemblée nationale, encodé et annoté de façon cohérente ; (clarin.eu)
- plus largement, les corpus **ParlaMint**, qui proposent des données parlementaires harmonisées pour plusieurs parlements européens, incluant la France, avec métadonnées riches et annotations syntaxiques/NER. (PMC)

Ces ressources sont très utiles pour l'étude du discours institutionnel (délibération, gouvernement/opposition, etc.), mais elles se concentrent sur un type de discours particulier (parlement) et n'intègrent généralement pas une annotation explicite d'**idéologie globale** au sens "droite / gauche / extrême, etc." : l'information passe surtout par les métadonnées de parti, de groupe, ou par le contexte historique.

1.2.2. Corpus de tweets et campagnes électorales

Une deuxième famille de ressources porte sur le **discours politique sur Twitter**, avec une orientation plus forte vers l'opinion et les campagnes :

- le corpus **Polititweets**, diffusé via la banque de corpus CoMeRe (ORTOLANG), rassemble plus de 34 000 tweets issus d'environ 205 comptes politiques influents (personnalités et comptes de partis). (repository.ortolang.fr)
- le corpus **#Présidentielle2017**, développé dans le cadre du projet **#Idéo2017**, stocke l'ensemble des tweets publiés par les candidats et, plus largement, autour de la campagne présidentielle de 2017, avec une infrastructure d'exploration basée sur Elasticsearch. (repository.ortolang.fr)
- des projets comme **Politoscope** ou d'autres observatoires sociologiques utilisent également des flux Twitter massifs pour analyser dynamiques de communautés, polarisation, circulation de hashtags, etc. (Cairn.info)

Ces corpus se prêtent bien à des tâches de **sentiment analysis** ou de **stance detection** : plusieurs travaux ont ainsi constitué des corpus annotés en positionnement (pour/contre, pro/anti) sur certains enjeux ou acteurs, par exemple un **corpus de 600+ tweets politiques français annotés pour la stance**, ou des threads annotés pour la prise de position dans une perspective plus conversationnelle. (ACL Anthology)

1.2.3. Corpus et vocabulaires politiques "classiques"

Enfin, de nombreux travaux s'appuient sur des corpus de :

- **presse politique** (éditoriaux, tribunes, dossiers thématiques) ;
- **professions de foi** et tracts électoraux (par ex. corpus de professions de foi parlementaires sur plusieurs décennies) ;
- **discours présidentiels** (Chirac, Sarkozy, Hollande, Macron, etc.), souvent constitués ad hoc pour un projet donné. (OpenEdition Journals)

Ces corpus sont généralement bien décrits dans les articles (taille, période, genres, méthodes), mais **ne sont pas toujours disponibles** comme ressources ré-utilisables standardisées (TEI, métadonnées détaillées, licence claire). Ils alimentent par contre une réflexion méthodologique sur :

- le **vocabulaire politique** : définitions, frontières, relations avec le social ; (Persee)
- la **variabilité sémantique** (singulier/pluriel, abstraction/concret, etc.) et ses liens avec les positionnements "de droite/de gauche" sur certains lexèmes. (perso.liris.cnrs.fr)

1.3. TAL / NLP et classification idéologique

Sur le versant TAL, deux tendances se dégagent :

1. **Corpus-assisted discourse studies / topic modeling** :
    - modélisation de thèmes dans les débats parlementaires (LDA, etc.), pour suivre l'importance relative de l'armée, de l'économie, de l'immigration, etc. sur le temps long ;
    - analyses assistées par corpus des champs lexicaux associés à des objets comme la "haine en ligne", les "fake news", la "liberté d'expression", etc. (OpenEdition Journals)
2. **Classification supervisée (stance, opinion, polarisation)** :
    - constitution de **corpus annotés manuellement** pour la stance en français, sur des enjeux politiques précis (référendums, scandales, élections), avec des étiquettes de type "pour/contre/neutre" ou "favorable/défavorable"; (ACL Anthology)
    - outils et plateformes comme **#Idéo2017**, qui combinent collecte, indexation, visualisation, mais où l'annotation idéologique reste souvent centrée sur les candidats ou les hashtags, plus que sur une typologie fine "droite / centre / gauche / extrêmes". (editions-rnti.fr)

Ces travaux montrent qu'il est possible d'entraîner des modèles performants pour reconnaître un **positionnement local** (pour ou contre une mesure, favorable ou hostile à un candidat), mais ils ne fournissent pas encore de **grand corpus français unifié**, couvrant plusieurs genres de discours longs, **explicitement annoté en "idéologie globale"** de manière homogène et réutilisable.

---

2. Positionnement de ton projet

2.1. Objectif général

Ton projet se situe à l'intersection :

- de cette tradition française d'**analyse des langages du politique** et de **lexicométrie** ;
- et des approches **NLP modernes** (spaCy, sklearn, Transformers HF) pour la **classification supervisée**.

Techniquement, le pipeline V5.5 repose sur :

- un **cœur générique** `core_prepare.py` / `core_train.py` / `core_evaluate.py` qui part de TEI XML, applique des mappings idéologiques définis en YAML, puis entraîne et évalue plusieurs familles de modèles (spaCy, sklearn, HF) sur des vues différentes (`ideology_global`, `left_intra`, `right_intra`, etc.) ;
- une couche d'**orchestration expérimentale** `superior_orchestrator.py`, qui lit des configs d'expériences, génère des grilles de runs (axes : dataset, stratégie d'équilibrage, famille de modèles, etc.), contrôle le parallélisme et la RAM, et produit des rapports agrégés.

Du point de vue des données, tu construis :

- un (**ou plusieurs**) **corpus TEI** volumineux (`web1`, `asr1`, éventuellement d'autres), stocké dans `data/raw/<corpus_id>/corpus.xml`, avec une structuration compatible avec ton pipeline (documents, métadonnées, acteurs, sources) ;
- un **référentiel d'idéologie** en YAML (`ideology.yml`), dérivé en plusieurs vues (`ideology_global.yml`, `ideology_left_intra.yml`, etc.), et appliqué automatiquement à chaque document via un mapping "acteur → idéologie".

Les premiers résultats montrent que, sur ta vue binaire `left/right`, des modèles comme un SVM TF-IDF atteignent déjà **environ 95 % d'accuracy** et un **macro-F1 autour de 0,84–0,85** sur un split `train/job` équilibré, avec un rappel très élevé pour la classe majoritaire et des performances raisonnables sur la minoritaire.

2.2. Ce que ton projet apporte par rapport à l'existant

Par rapport aux corpus et travaux existants, ton projet se distingue sur plusieurs plans :

1. **Type de corpus et diversité des sources**
    - Tu ne te limites ni au parlement (TAPS, ParlaMint) ni à Twitter (Polititweets, #Présidentielle2017), mais tu vises un **corpus multi-source** (web, éventuellement ASR, autres sites) avec des **textes longs** (articles, billets, tribunes, etc.). (clarin.eu)
2. **Annotation idéologique systématique**

- Ton **unité d'annotation** est l'**acteur / média** (parti, journal, organisation), décrit dans `ideology_actors.yml` et relié à une vue conceptuelle `ideology.yml`.
- Cette approche te permet de projeter automatiquement des labels idéologiques stables ("droite", "gauche", "extrême droite", etc.) sur des volumes massifs de textes, ce qui manque aux ressources existantes où l'idéologie est plutôt implicite (groupe parlementaire, hashtag, candidat).

3. **Cadre expérimental robuste**

   - Le pipeline est conçu pour **tester systématiquement** l'impact :

     - des stratégies d'équilibrage (`oversample`, `class_weights`, `cap_docs`, etc.),
     - des familles de modèles (`check`, `sklearn`, `spacy`, `hf`),
     - des découpages de corpus (mono-corpus, multi-corpus juxtaposé, cross-dataset).
   - L'orchestrateur `superior` permet de documenter précisément chaque run (`runs.tsv`, `metrics_global.tsv`, plots, `report.md`), ce qui rapproche ton projet d'un **banc d'essai expérimental** pour la classification idéologique en français.

4. **Standardisation TEI + exploitation NLP**

   - Tu assumes la contrainte TEI héritée des corpus parlementaires et des pratiques d'analyse du discours, mais tu l'intègres dans une chaîne contemporaine (spaCy, HF, sklearn), avec des formats intermédiaires standard (TSV, JSONL, DocBin).

En résumé, ton projet ne duplique pas un corpus pré-existant : il **combine** des choix de structuration (TEI, multi-corpus), d'annotation (idéologie globale multi-vues) et d'outillage (pipeline multi-familles + orchestrateur) qui n'existent pas, à ce jour, dans un ensemble français clé en main.

---

3. Pourquoi avoir *scrapé* un corpus (au-delà du plaisir de scraper)

On t'a explicitement demandé : "pourquoi avoir construit ton propre corpus au lieu de réutiliser un corpus existant ?". Tu peux répondre en trois temps.

3.1. Limites des corpus existants pour ta question

Les corpus existants sont précieux, mais ils ont chacun des limites pour **ta** question spécifique ("classification idéologique globale à partir de textes longs, multi-sources") :

- les corpus parlementaires sont **institutionnels** (un seul genre, un seul type de locuteur, format très normé) ; (Persee)
- les corpus Twitter sont **courts**, fortement contraints par la plateforme, et souvent centrés sur des campagnes électorales spécifiques, avec des catégories d'annotation plutôt "stance/opinion locale" qu'"idéologie globale". (ACL Anthology)
- les corpus de presse ou de professions de foi utilisés dans la littérature sont rarement publiés sous forme de ressources standardisées, et l'annotation idéologique est implicite. (OpenEdition Journals)

Autrement dit : **aucun corpus francophone existant ne fournit directement** ce dont ton pipeline a besoin :

- textes longs,
- multi-sources web,
- structuration TEI homogène,
- **annotation idéologique globale** cohérente,
- droits d'usage compatibles avec un usage de recherche + diffusion limitée.

D'où le choix raisonnable de **construire ton propre corpus**, plutôt que de bricoler des réutilisations partielles.

3.2. Intérêt scientifique et méthodologique du scraping

Le scraping n'est pas seulement un "plaisir coupable", c'est aussi un **choix méthodologique** :

1. **Aligner le corpus sur la question de recherche**

   - Tu peux cibler **les acteurs/médias** dont tu as besoin pour ton mapping idéologique (liste d'acteurs dans `ideology_actors.yml`), en t'assurant de couvrir différentes familles politiques de manière contrôlée.
   - Tu maîtrises la **période**, les genres, les formats, ce qui est crucial pour éviter les biais liés à la sur-représentation d'un événement ou d'un canal particulier (par ex. Twitter en période électorale).

2. **Contrôler l'hétérogénéité**

   - En partant d'un plan de scraping explicite (sites cibles, types de pages, critères d'inclusion), tu peux **documenter l'hétérogénéité** de ton corpus (médias, partis, supports) et compatibiliser cette hétérogénéité avec les recommandations de la textométrie. (perso.liris.cnrs.fr)

3. **Standardiser la structure (TEI) dès la collecte**

   - En imposant dès le scraping une structuration TEI minimalement cohérente (texte principal, métadonnées, balisage des acteurs,

dates, etc.), tu facilites la suite de la chaîne (TEI → TSV → formats modèles) et garantis la reproductibilité de tes expériences.

4. **Constituer une ressource potentiellement partageable**
    - Si les questions juridiques et éthiques sont correctement traitées (respect des conditions d'utilisation, anonymisation éventuelle, diffusion partielle ou contrôlée), tu peux, à terme, proposer une **ressource de référence** pour la communauté, qui manque aujourd'hui d'un corpus politique français "prêt à l'emploi" pour la classification idéologique.

3.3. Enjeux juridiques et éthiques (à connecter à ton mémoire d'épistémologie)

Enfin, le choix de scraper s'inscrit aussi dans le cadre plus large de ton mémoire "Les limites des machines : regards croisés culturels & juridiques" :

- le **statut juridique** des corpus web (text and data mining, exceptions de fouille de textes, conditions d'utilisation des sites, RGPD) pose des limites concrètes à ce que les chercheurs peuvent faire ;
- d'un point de vue éthique, la construction d'un corpus idéologique pose la question de la **représentation** (quels acteurs sont inclus / exclus), des risques de **stigmatisation** ou de **sur-exposition** de certains groupes, et de la manière dont on documente ces choix.

Tu peux donc justifier ton scraping en montrant que :

1. tu as **conscience** de ces limites et que tu les intègres à la réflexion (c'est un objet de ton mémoire) ;
2. tu ne scrapes pas "pour le fun", mais parce que c'est la **seule manière réaliste** d'obtenir un corpus adapté à ta question de recherche, tout en rendant visibles et discutables les choix de constitution.

---

4. Références commentées "à lire"

Voici une petite biblio de base, avec 1–2 phrases pour chaque :

Langages du politique / lexicométrie

- **Rémi-Giraud, S. (2010).** *Sémantique lexicale et langages du politique. Le paradoxe d'un mariage difficile ?* In *Mots. Les langages du politique*, 94. Synthèse sur ce que la sémantique lexicale peut (et ne peut pas) faire pour l'étude du politique ; très utile pour formuler la tension entre champs lexicaux et discours. (OpenEdition Journals)
- **Dubois, J. (1969).** *Le vocabulaire politique et social en France de 1869 à 1872.* Travail fondateur sur le vocabulaire politique, souvent cité comme point de départ des études lexicométriques du politique en France. (OpenEdition Books)
- **Tournier, M. (1982).** *Les vocabulaires politiques à l'étude, aujourd'hui (1962-...). Revue d'histoire moderne et contemporaine*, 62. Panorama des recherches sur les vocabulaires politiques, leurs méthodes et difficultés. (Persee)
- **Mayaffre, D. (2007).** *Les corpus politiques : objet, méthode et contenu. Corpus* (OpenEdition). Article très utile pour réfléchir à ce qu'est un "corpus politique", comment le constituer et le documenter. (OpenEdition Journals)
- **Pincemin, B. (2023).** *Hétérogénéité des corpus et textométrie. Langages*, 187. Discussion approfondie de l'hétérogénéité des corpus et de ses implications méthodologiques, avec des exemples sur le vocabulaire politique et syndical. (perso.liris.cnrs.fr)

Corpus parlementaires et TEI

- **Heiden, S. (1999).** *Encodage uniforme et normalisé de corpus. Application à l'étude d'un débat parlementaire. Mots. Les langages du politique*, 60. Pose les bases de l'encodage normalisé (SGML/TEI) pour les débats parlementaires, préfiguration des corpus actuels. (Persee)
- **TAPS-fr** – Transcription and Annotation of Parliamentary Speech (présenté dans des workshops CLARIN / LREC). Exemple de corpus parlementaire français encodé en TEI, pensé comme corpus monitor. (clarin.eu)
- **Erjavec, T. et al. (2022).** *The ParlaMint corpora of parliamentary proceedings.* Présentation des corpus ParlaMint, dont les débats parlementaires français, avec encodage harmonisé et annotations linguistiques. (PMC)

Corpus Twitter politiques et stance

- **Longhi, J. et al. (2014).** *Polititweets, corpus de tweets provenant de comptes politiques influents.* Banque de corpus CoMeRe (ORTOLANG). Description du corpus Polititweets, tweets de personnalités et comptes influents, exploité ensuite dans de nombreux travaux sur le tweet politique. (repository.ortolang.fr)
- **Longhi, J. (2013).** *Essai de caractérisation du tweet politique. Le discours et la langue*, 136. Réflexion sur les spécificités du tweet politique comme genre, avant même les grands corpus automatiques. (Persee)
- **Longhi, J. (2017).** *#Idéo2017 : corpus des tweets de la #présidentielle2017.* Doc. d'acquisition + articles associés. Décrit la collecte et l'exploitation des tweets de la présidentielle 2017, avec la plateforme #Idéo2017. (repository.ortolang.fr)
- **Evrard, M. et al. (2020).** *French Tweet Corpus for Automatic Stance Detection. LREC 2020.* Corpus annoté pour la stance (prise de position) sur

des tweets français, très proche de ce que tu fais côté modèle mais sur un autre type de données. (ACL Anthology)

- **Uro, R. et al. (2019).** *Création d'un corpus de tweets en français pour la détection automatique de stance.* Autre exemple de corpus de stance en français, plus petit mais méthodologiquement détaillé. (dahlia.egc.asso.fr)
