**Objectif général du projet** [**Discursive-telotopic-signatures**](https://github.com/Alan-Kleden/Discursive-telotopic-signatures)

Vérifier si des textes publics d’acteurs (partis, ministères, organisations…) contiennent des **signatures régulières** liées à leurs objectifs déclarés (“telos”), et si ces signatures aident à **prédire leurs positions** et leurs **évolutions dans le temps** mieux que des méthodes classiques.

**Vue d’ensemble par étapes**

**0) Préparer le terrain (1–2 jours)**

**But :** tout ranger pour que le travail soit clair et reproductible.

* **Créer le dossier de code** et l’organisation de base (collecte, calcul des variables, modèles, évaluation).
* **Verrouiller l’environnement logiciel** (fichiers requirements.txt, environment.yml) pour que tout le monde ait les mêmes versions.
* **Mettre un dépôt Git** (GitHub) pour suivre les changements.
* **Résultat attendu :** un projet qui démarre sans erreur sur n’importe quelle machine.

**1) Définir précisément les données (2–3 jours)**

**But :** savoir exactement **quels textes** on prend et **comment** on les identifie.

* **Qui ?** La liste d’acteurs (au moins 10, cible 15–20).
* **Quoi ?** Types de textes acceptés : discours, communiqués, tribunes, rapports officiels. Pas de posts courts ni de contenus personnels.
* **Quand ?** Deux périodes :
  + **T1 = entraînement** (par exemple M-24 à M-6)
  + **T2 = test** (M-6 à M) — jamais utilisée pour apprendre.
* **Où ?** Domaines de politique (au moins 2, idéalement 4 : Climat, Sécurité, Économie, Immigration).
* **Chocs externes** (événements marquants) : au moins 3 par domaine, datés à l’avance.
* **Résultat attendu :** trois fichiers propres et complets : actor\_roster.csv, shocks.csv, appendices A1–A7 (définitions et règles).

**2) Extraire le “telos” de chaque acteur (3–5 jours)**

**But :** résumer **les objectifs déclarés** d’un acteur en 1 à 3 phrases-cibles stables.

* **Sources :** pages “mission/charte/à propos”, programmes officiels.
* **Méthode simple :** on rassemble ces phrases, on regroupe celles qui se ressemblent, on garde 1 à 3 idées centrales.
* **Contrôle qualité :** 3 personnes annotent ; on mesure l’accord. Si l’accord est trop faible, on revoit la méthode.  
  **Seuil à atteindre :** indicateur d’accord ≥ 0,70.
* **Résultat attendu :** pour chaque acteur, un petit fichier avec ses “teloi” figés pour la suite.

**3) Construire les variables (features) à partir des textes (1–2 semaines)**

**But :** transformer les textes en **valeurs numériques** faciles à analyser.

* **Fc / Fi :** notes simples disant si un passage **promouvait** quelque chose (Fc) ou **s’opposait** (Fi).
* **θ (thêta) :** mesure d’**alignement** du texte avec le telos de l’acteur.  
  On en sort **cos(θ)** et **sin(θ)** pour avoir des nombres stables.
* **N\_tel :** **cohérence directionnelle** globale d’un texte (entre 0 et 1 environ).  
  Plus c’est haut, plus le document “tire” dans la même direction que le telos.  
  Les **poids** de calcul sont fixés à l’avance (pas de triche).
* **Ambivalence :** indicateur 1/0 si, dans une fenêtre de temps, on observe **à la fois** beaucoup de pro (Fc) et beaucoup de contre (Fi).
* **Fenêtrage temporel :** on regroupe par **fenêtres de 30 jours** en glissant de 7 jours, pour suivre l’évolution.
* **Résultat attendu :** deux tables :
  + **au niveau document** (une ligne par texte),
  + **au niveau fenêtre temporelle** (une ligne par période glissante).

**4) Établir les bases de comparaison (baselines) (3–5 jours)**

**But :** prouver que nos variables font mieux que des méthodes simples ou courantes.

* **Style-only :** prédire uniquement avec la forme du texte (longueur, ponctuation, etc.).
* **BERT-stance :** modèle standard qui prédit “pour/contre” sans nos variables telos.
* **Party-line prior :** une règle simple basée sur l’historique T1.
* **Résultat attendu :** code qui produit les scores de ces baselines sur T2.

**5) Faire les tests principaux (H1, H2, H3) (1–2 semaines)**

**But :** vérifier les **hypothèses** annoncées dans la préinscription.

* **H1 :**  
  Ajouter θ et N\_tel à un modèle déjà basé sur valeurs, valence, Fc/Fi **améliore** la prédiction sur T2.  
  **Mesure :** AUC (aire sous la courbe ROC).  
  **Succès :** amélioration ≥ 0,05 et intervalle de confiance qui n’englobe pas 0.
* **H2 :**  
  Entraîner sur un domaine A (ex. Climat) et tester sur un domaine B (ex. Sécurité).  
  **Succès :** notre modèle ≥ modèle BERT-stance + 0,05 d’AUC sur T2.
* **H3 :**  
  Sur les séries temporelles, nos variables expliquent **plus de variance** autour des chocs (ambivalence, re-cohérence).  
  **Succès :** gain d’explication (ΔR²) ≥ 0,05, test statistique positif.
* **Résultat attendu :** un tableau clair de résultats avec intervalles de confiance, corrections des multiples tests, et un verdict “réussi / partiel / non réussi” par hypothèse.

**6) Transparence et reproductibilité (en continu)**

**But :** que n’importe qui puisse **refaire** l’étude et **vérifier**.

* **Seeds fixés :** mêmes nombres aléatoires → mêmes résultats.
* **Données figées :** on fige T1/T2 **avant** d’analyser.
* **Code public :** dépôt Git avec instructions claires (README).
* **Environnement :** fichiers pour recréer les mêmes versions de logiciels.
* **Résultat attendu :** un dossier **OSF** et un **GitHub** où tout est expliqué et rejouable.

**Gouvernance simple (qui fait quoi)**

* **Lead méthode/données :** valide la liste d’acteurs, les fenêtres, et le respect des règles.
* **Lead extraction telos :** pilote l’annotation et l’accord entre annotateurs.
* **Lead features/modèles :** implémente les calculs et les tests H1/H2/H3.
* **Repro & infra :** s’assure que tout tourne partout (env, scripts, documentation).

**Critères de “go / no-go”**

* **Go :** telos extraits avec accord suffisant (≥ 0,70), au moins 10 acteurs, chocs complets, tables de features stables.
* **No-go :** si telos de qualité insuffisante pour plus de 50 % des acteurs, on **documente l’échec** et on s’arrête (comme promis dans la préinscription).

**Risques principaux et parades**

* **Qualité du telos insuffisante :** faire un **pilote** sur 3–5 acteurs, corriger les consignes d’annotation.
* **Données trop faibles dans un domaine :** élargir légèrement la période ou choisir des acteurs plus prolifiques **avant** le gel.
* **Temps de calcul long :** tester d’abord sur **un sous-ensemble** pour valider la chaîne, puis passer en complet.

**Livrables finaux (clairs et utiles)**

1. **Jeu de résultats** H1, H2, H3 sur T2, avec intervalles et décisions.
2. **Rapport** court : ce qui marche, ce qui ne marche pas, limites, pistes.
3. **Code + environnement** rejouables, documentés.
4. **Annexes** (définitions, poids, règles) alignées avec l’OSF.

**Glossaire minimal**

* **Telos :** objectifs déclarés d’un acteur (ce qu’il dit vouloir atteindre).
* **θ (thêta) :** angle d’alignement entre un texte et le telos (converti en nombres cos/sin).
* **N\_tel :** mesure de cohérence globale d’un document avec le telos (plus haut = plus cohérent).
* **AUC :** score de qualité de prédiction (0,5 = hasard, 1,0 = parfait).
* **ΔR² :** gain d’explication dans un modèle temporel.