Responsabilités et injonctions dans un corpus de réponses citoyennes sur l'écologie: analyse lexicométrique sous TXM

Belosevich Anastasiia (a.belosevich@gmail.com ; numéro étudiant 21701822),
Martin Marina-Bella (martin.marina-bella@gmail.com ; numéro étudiant 22404707)

I. Introduction

(Martin Marina-Bella)

L'écologie est au cœur des discussions publiques; elles abordent les menaces qui pèsent sur l'environnement et, d'un autre côté, la manière de mieux le protéger.

Le corpus analysé dans le cadre de ce travail est issu du Grand Débat National organisé par l'État français, où les citoyens débattent de questions essentielles liées à la transition écologique.

L'analyse vise à identifier les différents acteurs désignés comme responsables, mais également les actions mises en avant pour faire face à ces enjeux.

Afin de répondre à cette question, nous utiliserons le logiciel TXM, qui nous permettra d'analyser les fréquences lexicales, les concordances et les cooccurrences. Dans un premier temps, nous analyserons les responsables de la transition écologique. Dans un second temps, nous nous intéresserons à la formulation des actions pour la protection de l'environnement.

Ci-dessous sont présentées les problématiques que nous avons retenues ainsi que les hypothèses qui y sont associées.

Problématique 1:
à qui les répondants attribuent-ils la responsabilité de la transition écologique?

Hypothèse 1:
la responsabilité écologique est majoritairement attribuée à des acteurs institutionnels (État, gouvernement, entreprises), tandis que le citoyen est principalement responsabilisé sur le plan moral et comportemental.

Problématique 2:
quelles actions peuvent être engendrées pour sauver la planète ?

Hypothèse 2:
les actions pour préserver l'environnement sont principalement exprimées par des verbes d'action, tandis que l'usage du terme "faudrait" vise à formuler des recommandations plutôt qu'une mise en œuvre immédiate.

Toutes les captures d'écran sont présentées dans le fichier *images.pdf*.

Les scripts Python et les fichiers CSV, ainsi que les corpus au format XML, mentionnés dans ce travail, se trouvent dans le dépôt Git accessible à l'adresse suivante:
github.com/abelosev/extraction_2025

II. Préparation du corpus

(Belosevich Anastasiia)

2.1 Transformation du format

2.1.1 Une brève description du corpus initial

Le corpus analysé est:

- argumentatif,

- composé de nombreuses phrases longues avec une ponctuation complexe,

- structuré de manière telle que chaque ligne correspond à une réponse complète, sans ligne vide pour séparer les réponses,

- partiellement rédigé sous forme de listes.

On y trouve plusieurs caractères utilisés pour marquer les éléments d'une liste : « • », « ● », « * », ainsi que 3 types de tirets: trait d'union, tiret moyen (en dash) et tiret long (em dash). Il est probable que les éléments de liste appartenant à une même réponse se trouvent souvent sur une seule ligne, mais par précaution, nous avons également prévu le cas où ils pourraient être répartis sur plusieurs lignes.

2.1.2 Pourquoi structurer le corpus ?

Le corpus dans son état initial (un grand fichier TXT non structuré) n'a pas pu être importé dans TXM. Il a donc fallu le convertir au format TEI-XML afin de permettre son importation et son traitement. Cela permet de réaliser des analyses plus approfondies, en plus le corpus devient plus lisible et plus clair.

2.1.3 La structuration du corpus

À l'aide du script Python **make_tei.py**, le document TXT a été converti au format TEI-XML.
La structure générale est la suivante:

```
<TEI>
 <teiHeader> ... </teiHeader>
 <text xml:id="corpus">
  <body>
   <div type="response" xml:id="resp000XX" n="XX" words="…">
    <p>…</p>
   </div>
  </body>
 </text>
</TEI>
```

Figure 1 du fichier **images.pdf** représente un exemple de réponse contenant une liste. On voit que toutes les phrases sont regroupées dans un seul paragraphe.

Le passage au format TEI permet de structurer clairement le corpus pour son exploitation dans TXM, en traitant chaque réponse comme une unité d'analyse distincte et en lui associant des métadonnées utiles, ce qui facilite les analyses comparatives.

Nous allons analyser un bloc du document TEI résultant:

```
<div type="response" xml:id="resp00001" n="1" words="4">
  <p>Multiplier les centrales géothermiques.</p>
</div>
```

Chaque réponse est encodée dans un élément , qui correspond à l'unité d'analyse. Les attributs **xml:id** et **n** servent à identifier et ordonner les réponses, tandis que **words** indique leur longueur et permet des comparaisons quantitatives.

Sur les captures d'écran suivantes (*voir Figures 2, 3*), on peut voir la fenêtre de TXM avec le début et la fin du corpus importé (le début et la fin correspondent au fichier TXT d'origine): on constate que, grâce au format TEI, les réponses sont clairement séparées et que le corpus est beaucoup plus lisible.

2.2 Corpus complémentaire

Initialement, il était prévu de réaliser l'analyse exclusivement à l'aide de TXM. Toutefois, dans la pratique, l'utilisation de TXM s'est révélée partiellement limitée: dans notre configuration, il n'a pas été possible d'activer la lemmatisation, le module *Calculette* ni l'intégration avec *R*, et il était également impossible d'exclure correctement les stop-words.
Dans ce cadre, il a été décidé de créer une version supplémentaire du corpus, nettoyée à l'aide de la bibliothèque spaCy. Cette version est utilisée pour des analyses quantitatives complémentaires, tandis que l'analyse principale (concordances, étude des contextes et interprétation discursive) reste fondée sur le corpus original dans TXM.

Corpus nettoyé (*voir Figures 4*).

2.3 Unités d'analyse

Nous allons considérer deux niveaux d'unités pour l'analyse: le niveau des réponses (*div*) et le niveau des tokens.

Figures 5 présente une capture d'écran des Corpus properties. On voit que le corpus

comprend près de 10 millions de mots. On constate également que le corpus est structuré en plusieurs niveaux (*text, body, div, p*), ce qui permet de définir les réponses comme unités d'analyse au niveau des sections.

Malheureusement, les outils disponibles dans TXM (version 0.8.4) n'ont pas permis de calculer les longueurs minimale, maximale et moyenne des réponses du corpus. Ce problème a donc été résolu à l'aide du script Python **len_script.py**, qui a analysé le fichier XML d'origine. Les résultats obtenus sont présentés sur la *Figure 6*.

Ces données peuvent être utilisées pour constituer des sous-corpus de réponses développées et de réponses courtes de type slogan. Les réponses très courtes, comme dans l'exemple donné ci-dessus (*Multiplier les centrales géothermiques*), prennent souvent la forme d'énoncés injonctifs, sans mention explicite d'un acteur responsable, ce qui les rapproche davantage de slogans que d'argumentations développées.

Malheureusement, en raison du volume limité de ce travail, nous n'avons pas pu mener une analyse de ces sous-corpus. Toutefois, cela constitue une piste possible pour de futures recherches sur ce corpus.

III. Analyse de l'attribution de la responsabilité écologique

(Belosevich Anastasiia)

Dans cette section, nous utiliserons des méthodes textométriques afin d'analyser la première partie de la problématique formulée dans l'introduction.

Nous analyserons:
1/ les acteurs de la responsabilité,
2/ les marqueurs de l'obligation (*il faut, doit*),
3/ la polarité "individu vs système" et la repartition de la responsabilité

3.1 Étude des "acteurs" de la responsabilité

Dans cette partie, nous analysons de quels acteurs parlent les répondants lorsqu'ils évoquent la responsabilité écologique, sans tirer pour l'instant de conclusions sur l'attribution de cette responsabilité.
Nous avons utilisé un corpus nettoyé à l'aide de spaCy afin de produire une table de fréquences dans TXM (*voir Figures 7*).

Le résultat obtenu n'est pas vraiment représentatif, car il regroupe des formes relevant de différentes catégories grammaticales (je rappelle que, dans ma version de TXM, la lemmatisation ne fonctionnait pas). Nous pourrions examiner manuellement les 100 à 200 mots les plus fréquents afin d'établir une liste approximative d'acteurs, mais, d'une part, cela prendrait beaucoup de temps et, d'autre part, il y aurait un risque de passer à côté de termes moins fréquents mais plus pertinents (par exemple, des mots rares mais utilisés dans des contextes accusatoires).

L'approche suivante est donc proposée: effectuer une lemmatisation et une annotation morpho-syntaxique (POS) avec spaCy, extraire les NOUN/PROPN, filtrer selon un seuil minimal de fréquence (≥ 20), sélectionner manuellement 5-10 termes correspondant à la notion d'acteur, puis vérifier leur emploi dans TXM à l'aide des concordances.

Par "acteurs", on peut entendre plusieurs catégories: des personnes (y compris des noms propres) et des groupes (*citoyens, Français*), des institutions (*État, gouvernement*), des organisations (*entreprises, banques*), ainsi que des notions liées au système (*capitalisme*).

Nous regrouperons ces éléments en 3 ensembles: acteurs individuels (*citoyens*); acteurs institutionnels et organisationnels (*État, entreprise, gouvernement, lobby*); acteurs implicites ou abstraits (*système, société, industrie*). Limite de cette méthode est celle qu'elle laisse de côté des adjectifs comme *jeunes, de gauche*, etc.

Observons maintenant les 20 lemmes les plus fréquents obtenus avec le script

Python **filtrage_actors.py** (le tableau complet figure dans le fichier joint *actors.csv*): *voir Figure 8.*

L'analyse des lemmes les plus fréquents montre une forte présence du terme *entreprise* et, à l'inverse, une faible visibilité des acteurs individuels, ce qui suggère un déplacement de la responsabilité vers des acteurs institutionnels. L'absence de noms propres parmi les formes fréquentes confirme également que la responsabilité est peu personnalisée et construite avant tout comme un problème structurel et systémique, ce qui produit un effet de "responsabilité diffuse" que nous approfondirons à travers l'analyse des concordances.

Afin d'identifier des candidats pertinents pour l'analyse des concordances, nous avons d'abord examiné la liste des lemmes nominaux les plus fréquents (environ les 100 premiers). À partir de cette liste, un sous-ensemble d'environ 10 lemmes a été retenu pour une analyse qualitative approfondie.

1/ Acteurs individuels: *citoyen* (4768),

2/ Acteurs collectifs: *entreprise* (10419), *État* (5316) /*etat* (4654) /*gouvernement* (3578). Nous ajoutons également à cette liste le terme *lobby* (1975), afin de montrer que des mots peu fréquents peuvent néanmoins être importants pour l'analyse lorsqu'ils apparaissent dans des contextes pertinents.

3/ Acteurs implicites /abstraits: *société* (4059), *système* (3977).

Un point important est à souligner: spaCy a correctement lemmatisé l'ensemble des termes retenus, à l'exception du mot *lobby*. Cela s'explique par le fait qu'il s'agit d'un emprunt relativement récent dont l'orthographe varie. Pour cette raison, nous avons regroupé manuellement les formes suivantes: *lobie* (22), *lobbyiste* (202), *lobbyes* (9) et *lobby* (1742), ce qui conduit à une fréquence totale de 1975.

3.2 Analyse des concordances

Dans cette partie, nous nous sommes limités à l'étude des concordances pour **citoyen** et **État/gouvernement**. Les concordances ont été analysées de manière qualitative afin d'identifier les rôles discursifs attribués aux différents acteurs, en portant une attention particulière à la position grammaticale (sujet, objet), à l'expression de la responsabilité (obligation: *doit, devrait*; contrainte: *imposer, interdire*; accusation: *responsable, coupable*; déplacement de la faute) ainsi qu'à la polarité: responsabilité positive (*agir, protéger, investir*) et négative (*polluer, bloquer, profiter*).

3.2.1 Acteur individuel: citoyen

A. Position grammaticale

Dans les concordances, citoyen apparaît à la fois en position de sujet (*les citoyens doivent…, tout citoyen doit faire un effort*) et en position d'objet, ce qui le présente comme une cible d'injonctions ou de politiques publiques (*culpabiliser le citoyen, faire porter la responsabilité au citoyen*).

Cooccurences du lemme "citoyen" dans le corpus nettoyé: *voir Figure 9*.

Exemple de concordances pour le syntagme "citoyen doit": *voir Figure 10*.

B. Type de responsabilité attribuée

Les concordances (*voir Figure 11*) montrent que le citoyen est surtout présenté comme responsable sur le plan moral et comportemental, souvent soumis à des obligations (*il faut responsabiliser le citoyen*) et à des contraintes (*taxer les citoyens*, *demander aux citoyens de changer leurs comportements).*

Plusieurs concordances (*voir Figure 12*) mettent en évidence un discours critique de la culpabilisation: *arrêter de culpabiliser le citoyen*, *on fait porter la responsabilité uniquement sur le citoyen.*

Les locuteurs dénoncent explicitement un déséquilibre dans l'attribution des responsabilités: *arrêter de faire porter le chapeau au citoyen qui est un faible responsable en ce domaine*.

C. Polarité

Le discours construit une image ambivalente du citoyen. S'il est parfois présenté comme pollueur, cette idée apparaît surtout dans des contextes critiques où elle est contestée ou relativisée (*ce n'est pas le citoyen qui pollue le plus*).

Les concordances montrent ainsi que le citoyen est moins désigné comme véritable responsable écologique que comme cible d'une culpabilisation, souvent mise en contraste avec la responsabilité des entreprises ou des industries (*voir Figure 13*).

D. Conclusions

Ainsi, les concordances montrent que le citoyen occupe une place centrale mais ambivalente: il est fortement appelé à changer ses comportements, tout en étant présenté comme ayant peu de pouvoir réel face à l'État ou aux entreprises, ce qui renforce une responsabilité diffuse et culpabilisante.

3.2.1 Acteur institutionnel: État/gouvernement

A. Position grammaticale

Dans le corpus, l'État et le gouvernement occupent majoritairement la position de sujet dans des constructions modales fortes (*l'État doit, le gouvernement impose*), ce qui les présente comme des acteurs centraux de l'action écologique, dotés d'initiative et de pouvoir décisionnel (*voir Figure 14*).

B. Type de responsabilité attribuée

Les contextes montrent que l'État et le gouvernement sont surtout présentés à travers des expressions d'obligation (*doit*, *il faut que*). Ils disposent également d'un pouvoir de contrainte (*imposer*, *interdire*, *taxer*), généralement considéré comme nécessaire mais insuffisant. Enfin, le discours critique souvent l'inaction de l'État et dénonce explicitement une stratégie de déplacement de la faute (*voir Figure 15*), attribuée aux institutions elles-mêmes (*on demande trop aux citoyens alors que l'État…*; *l'État ne peut pas faire porter la transition sur les citoyens*).

C. Polarité

Quand l'État agit, la polarité est très positive (*agir*, *protéger*, *financer*, *investir*, *planifier*) comme il est perçu comme le seul acteur capable d'agir à l'échelle systémique. En même temps, le gouvernement est souvent critiqué pour l'inaction ou l'incohérence (*laisser faire*, *ne pas réguler*, *maintenir des politiques polluantes*, *être complice des lobbies* - voir Figure 16).

D. Conclusions

Contrairement au citoyen, dont la responsabilité est souvent contestée ou relativisée, celle de l'État apparaît comme structurelle, centrale et non substituable, ce qui renforce l'idée d'un discours orienté vers une responsabilité systémique et institutionnelle.

Le cadre de ce travail ne nous permet pas d'analyser de la même manière les contextes associés à *entreprise* et *lobby*, toutefois cette démarche constituerait une étape logique pour approfondir l'analyse de la problématique.

Passons maintenant à l'analyse des marqueurs d'obligation afin de confirmer ou d'infirmer l'hypothèse sur la répartition de la responsabilité entre les individus et l'État.

3.3 Analyse des marqueurs d'obligation

À l'aide de spaCy, nous établissons la distribution fréquentielle des verbes dans le corpus (script Python **filtrage_verbs.py,**

les résultats sont présentés dans le fichier **verbs.csv**):

```
lemma,frequency
falloir,37984
mettre,19780
faire,19155
développer,13562
produire,12750
pouvoir,12633
devoir,12497
arrêter,12187
taxer,11078
prendre,10631
favoriser,10559
permettre,10473
interdire,9310
```

Voici quelques marqueurs d'obligation trouvés parmi les lemmes les plus fréquents: *falloir* (surtout "il faut que"), *devoir*, *arrêter*, *interdire*, *obliger* (fréquence 4534).

Pour analyser les destinataires typiques des exigences formulées par les répondants, nous nous limiterons à l'analyse des cooccurrences et des concordances de "il faut que" et de "doit / doivent".

Les cooccurrences montrent une forte centralité de l'État comme sujet grammatical des verbes modaux forts (*il faut que l'État…),* ce qui nous conduit à conclure que l'État est construit comme le principal responsable et comme le détenteur de la capacité d'action (*voir Figures 17, 18*).

Des résultats similaires sont obtenus lors de l'analyse des cooccurrences de *doit* / *doivent*:

Query doit

Cooccurrent	Frequency	CoFrequency	Score	Mean distance
priorité	2718	601	297	2.8
Etat	4836	764	274	1.4
état	6940	919	270	1.6
passer	3421	623	258	.8

Il est important de noter que l'utilisation d'un corpus nettoyé des pronoms (intégrés aux stop-words) pour l'identification des acteurs a conduit à une sous-représentation de formes telles que *chacun, tout le monde, chaque Français* ou *chaque citoyen*. Or, ces formes jouent un rôle central dans la construction discursive de l'obligation, en fonctionnant comme des acteurs individualisés mais anonymes, porteurs d'une injonction généralisée (*voir Figure 19*).

3.4 Conclusions

Les cooccurrences et concordances étudiées montrent que les mesures écologiques sont majoritairement formulées sous la forme d'injonctions normatives. L'État et le gouvernement apparaissent comme les acteurs centraux de la contrainte et de la mise en œuvre, tandis que les citoyens sont principalement interpellés à travers des obligations morales et comportementales (ce qui a été montré dans la partie 3.2.1), souvent formulées via des acteurs diffus (*chacun, tout le monde*). On peut ainsi considérer que l'hypothèse formulée dans l'introduction est confirmée.

IV. Problématique 2 : quelles actions peuvent être engendrées pour sauver la planète ?

(Martin Marina-Bella)

Hypothèse : les actions pour préserver l'environnement sont majoritairement exprimées à travers des verbes d'action, et l'usage du mot "faudrait" a pour but d'indiquer et de mettre en place des conseils à suivre, et non une mise en œuvre immédiate.

Pour cette seconde partie, nous avons fait le choix de nous concentrer sur les verbes d'action car ils interprètent la volonté des citoyens et montrent le dynamisme du changement pour protéger l'environnement. L'analyse s'appuiera sur les fréquences et les concordances.

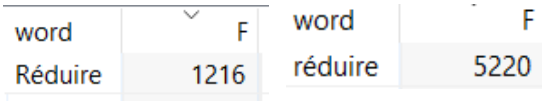
Nous nous intéressons particulièrement aux verbes d'action tels que "réduire", "accompagner", "encourager", "favoriser" ainsi qu'au terme "faudrait" qui n'est pas un verbe d'action en soi mais une démarche pour conseiller une action.

4.1 Analyse de l'index

L'analyse de fréquence montre une présence assez importante des verbes

d’action. Le verbe réduire apparaît 6436 fois.

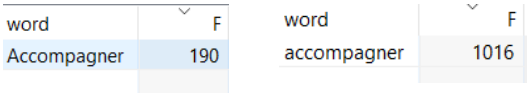
Exemple:



word	F
Réduire	1216

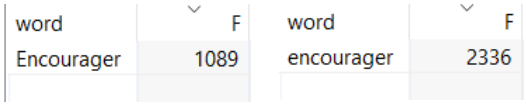
word	F
réduire	5220

Tandis que le mot accompagner possède 1406 et encourager totalise 3425 occurrences.



word	F
Accompagner	190

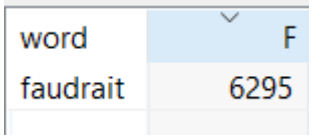
word	F
accompagner	1016



word	F
Encourager	1089

word	F
encourager	2336

Pour finir, le terme "faudrait" n’est pas utilisé en tant que verbe d’action, mais il a pour but d’énoncer une action et d’en faire une nécessité; il comptabilise un total de 6295 occurrences.



word	F
faudrait	6295

Les fréquences de chaque verbe d’action sont assez élevées et orientent le discours vers la volonté d’intervenir et de faire quelque chose; pour étudier plus en profondeur, il serait pertinent d’analyser les contextes des verbes grâce aux concordances. Cependant, la partie pivot droite nous intéressera un peu plus pour cette étude.

4.2 Analyse des concordances

Le verbe "réduire" a un total de 5220 occurrences, il est utilisé pour désigner la diminution de certains éléments perçus de manière négative, tels que l’impact écologique, les émissions de CO_2, la pollution, le carbone et les prix, ce qui exprime la volonté de préserver l’environnement (*voir Figure 20*).

Les verbes "accompagner" et "encourager" ont respectivement 1016 et 2336 occurrences; ils sont généralement utilisés pour désigner le soutien particulier qu’il convient d’apporter à la transition, aux projets, aux citoyens et aux entreprises. Cela souligne l’importance des acteurs et des projets pour l’environnement *(voir Figure 21)*.

Quant au terme "faudrait", il contient 6295 occurrences. Cela prouve une force de proposition, à travers des formulations telles que "faudrait avoir" et "faudrait faire"; le but est d’initier une action collective, "faire quelque chose ensemble pour le bien de l’environnement" (*voir Figure 22*).

L’ensemble des verbes d’action dans le corpus permet de mettre en évidence le fait que les citoyens veulent agir, soutenir, et le verbe "faudrait" permet d’introduire des actions à faire sans pour autant les dicter en obligation.

4.3 Etude de la progession des verbes

Grâce à la progression, nous pouvons observer que l’emploi des verbes ne cesse d’évoluer, ce qui traduit une forte volonté d’entreprendre des actions (*voir Figure 23*).

4.4 Conclusions

Généralement, les forces de proposition sont axées sur l’écologie et la réduction de certaines consommations: la pollution, le carbone, les énergies fossiles, le chauffage et les coûts.
L’accompagnement et l’encouragement des entreprises, des citoyens et de l’État, ainsi que la transition, sont assez ciblés; en ce qui concerne les actions telles que "organiser des débats", "utiliser moins d’énergie" et "promouvoir l’innovation", elles sont essentielles pour améliorer l’équilibre environnemental.

V. Conclusions et Perspectives
(Martin et Belosevich)

5.1 Conclusions

L'analyse confirme que la responsabilité écologique est principalement attribuée à l'État, au gouvernement et aux entreprises; les citoyens ont davantage une responsabilité sur le plan moral et comportemental.

En ce qui concerne les actions, les verbes indiquent une orientation visant à réduire des éléments négatifs, tels que la pollution, le carbone et le CO_2, tout en mettant en évidence l'importance de l'accompagnement nécessaire pour soutenir la protection de l'environnement.

L'usage fréquent du terme «faudrait» met en évidence les recommandations formulées par les citoyens et l'importance de la collectivité visant à inciter à l'action.

Le discours constitue un recueil de conseils visant la réflexion et l'incitation, plutôt qu'une approche directe et opérationnelle.

5.2 Perspectives

Dans la suite de cette analyse, il serait pertinent de constituer des sous-corpus distincts selon les acteurs de la responsabilité écologique (citoyen, État/gouvernement, entreprises/lobbies). Cette démarche permettrait de comparer plus précisément les discours associés à chaque acteur et d'observer comment se répartissent les marqueurs d'obligation, de contrainte ou d'accusation.

Une autre piste de recherche consisterait à analyser l'usage des pronoms personnels (*je*, *nous*, *on*), afin de mieux comprendre le positionnement des répondants par rapport à la responsabilité écologique. Cette analyse permettrait de distinguer les formes d'implication individuelle, d'identification collective et de mise à distance de l'action.

En combinant ces deux approches, il serait possible de mieux analyser la manière dont la responsabilité est formulée dans le corpus.