# Le dire ou l'écrire sur les réseaux sociaux numériques

Pierre Ratinaud, Brigitte Sebbah

# Le dire ou l'écrire sur les réseaux sociaux numériques : comparaison des messages écrits et des messages audios sur la chaîne Telegram des gilets jaunes de Haute-Garonne[1]

Pierre Ratinaud[1], Brigitte Sebbah[2]

[1]LERASS, Université de Toulouse Jean Jaurès – pierre.ratinaud@univ-tlse2.fr

[2]LERASS, Université de Toulouse Paul Sabatier – brigitte.sebbah@iut-tlse3.fr

## Abstract

Those involved in the Yellow Vests movement have used numerous communication channels, most of which come under the heading of socio-numerical networks (Facebook, Twitter, WhatsApp...). We gained access to the Telegram feed of the Haute-Garonne yellow vests and set about analyzing the discussions held there. Initially, we looked at the 50,000 written messages, seeking to determine the different themes addressed based on an analysis with the Reinert method (Sebbah & Ratinaud, 2023) in the IRaMuTeQ software (Ratinaud, 2020). The availability of the Whisper model (Radfort & al., 2022), which enables the automatic transcription of oral discourse, also enabled us to analyze the 2201 audio messages present on the discussion threads. After highlighting some of the limitations of the automatic transcription produced by Whisper, we will focus on the differences and similarities between text and audio messages. Our first remarks will concern the statistics associated with these corpora (average message size, word frequencies, etc.), then we'll compare the preferred themes.

**Keywords:** Telegram, audio transcritpion, Whisper, oral, writing

## Résumé

Les acteurs du mouvement des gilets jaunes ont utilisé de nombreux canaux de communications relevant majoritairement des réseaux socionumériques (Facebook, Twitter, WhatsApp...). Nous avons eu accès au fil Telegram des gilets jaunes de Haute-Garonne et avons entrepris l'analyse des discussions qui s'y sont tenues. Nous nous sommes dans un premier temps intéressés aux 50 000 messages écrits en cherchant à déterminer les différentes thématiques abordées à partir d'une analyse avec la méthode Reinert (Sebbah & Ratinaud, 2023) dans le logiciel IRaMuTeQ (Ratinaud, 2020). La mise à disposition du modèle Whisper (Radfort & al., 2022) qui permet la transcription automatique du discours oral nous a permis d'envisager également l'analyse des 2201 messages audios présents sur les fils de discussion. Après avoir souligné quelques limites de la retranscription automatique produite par Whisper, nous nous attacherons à pointer les différences et les similarités entre les messages textuels et les messages audios. Nos premières remarques concerneront les statistiques associées à ces corpus (taille moyenne des messages, fréquences des mots, etc...), puis nous comparerons les thèmes privilégiés.

**Mots clés :** Telegram, transcription de l'audio, Whisper, oral, écrit

## 1. Introduction

Décembre 2018, les Gilets Jaunes s'apprêtent à manifester lors de l'acte 3 en étant désormais structurés hors des réseaux socionumériques dits traditionnels. Si ce mouvement qui revendique une horizontalité, l'absence de représentant et donc de porte-parole, s'est constitué dès

[1] Ce travail a été réalisé dans le cadre du Labex SMS, portant la référence ANR-11-LABX-0066, a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'avenir portant la référence ANR-11-IDEX-0002-02.

novembre 2018 à la fois sur les ronds-points des villes françaises et déployé sur une myriade de pages *Facebook*, c'est sur une messagerie souterraine et réputée pour son anonymat et la sécurité de ses données qu'il va poursuivre son déploiement. C'est dans ce contexte que nous avons eu la possibilité d'explorer les fils de discussions d'un réseau social peu étudié, *Telegram*, en analysant les messages des canaux d'un groupe hyperlocal toulousain du 1er novembre 2018 (date de création du groupe MVT31) au 11 avril 2019 (date de fermeture du groupe)[2]. Notre premier objectif était de compléter les travaux sur les pages visibles et officielles des gilets jaunes (*Twitter*, *Facebook*) (Sebbah et al, 2019; Souillard et al, 2019; Marchand et al, 2019; Baisnée et al, 2019; Cointet et al., 2021; ) en décrivant le contenu des messages textuels produits. Mais la nature des échanges sur ce dispositif, qui mèle messages écrits et messages audios, offre également l'occasion d'explorer les différences entre communication écrite et communication orale sur les réseaux socionumériques. Notre première approche avez dû faire le deuil de l'étude de ces « voice messages », au même titre que celle des images et des vidéos, mais la mise à disposition du modèle Whisper (Radfort & al., 2022) qui permet la transcription automatique du discours oral, nous a finalement permis d'envisager l'analyse de ces 2201 enregistrements. Cette étude se centre donc sur la comparaison entre messages écrits et messages oraux sur Telegram. Après avoir rappelé le contexte de production de ces données, nous ferons un premier bilan des transcriptions proposées par Whisper, puis nous comparerons les productions écrites et les productions orales, d'abord du point de vue des statistiques lexicales, puis du point de vue des thématiques abordées dans ces messages.

## 2. Le Canal Telegram MVT31

Le canal des gilets jaunes toulousains était constitué de 7 fils de discussion. Le tableau 1 présente le nombre d'abonnés de chacun de ces fils :

| | abonnés |
|---|---|
| Info actions | 1030 |
| Annonces fiables | 867 |
| Bla,bla, échanges | 710 |
| Sondages | 418 |
| CR | 410 |
| Covoiturage | 255 |
| Justice et témoignages | 168 |

*Tableau 1 : Fréquence des abonnés par fil de discussion*

Le canal InfoAction est le premier à avoir été créé, 15 jours avant la première manifestation des gilets jaunes. Il compte 1030 utilisateurs qui ont généré 14539 messages. Le canal Blabla apparaît 4 jours après la première manifestation. Il a réuni 10 utilisateurs qui ont posté 3 1 0 messages. 'un point de vue méthodologique, l'analyse que nous proposons s'appuie sur les données extraites des fils la bla, échanges et InfoAction qui sont les seuls à avoir donné lieu à des discussions.

[2] Nous tenons à remercier Guillaume Cabanac pour l'aide apportée dans l'extraction de ces données

La création du fil Blabla est postérieure à celle du fil InfoAction. A partir du 30 novembre 2018, l'usage de ce fil InfoAction s'arrête complètement pour reprendre le 18 décembre avec un message de l'administrateur qui pour la première fois va rappeler la fonction de chacun des fils en précisant qu'il n'est pas responsable des actions ou des réunions organisées. Le fil ne retrouvera une activité normale qu'à partir du 2 décembre 2018. L'arrêt soudain des conversations coïncide avec une opération policière de grande ampleur qui a mis fin à l'occupation de la plupart des lieux de lutte des gilets jaunes. En outre, dès le lendemain, les manifestations vont faire entrer le mouvement dans une autre dimension en étant marquées par des épisodes de violence, notamment sur les Champs-Elysées. On recense 133 blessés ce jour-là parmi les manifestants, dont plusieurs éborgnés, des pillages et 249 incendies dans toute la France. Le fil blabla quant à lui, restera actif durant cette période et jusqu'à sa destruction.

## 3. Transcription des fichiers audio

### *3.1. Description et analyse des transcriptions*

Nous avons donc utilisé le modèle Whisper[3] (Radfort & al., 2022) proposé en open-source par la société OpenIA[4] pour retranscrire les 2201 messages audios non-vides laissés sur les fils Info Action et blabla. Les différents modèles disponibles sur la page du projet Whisper[5] permettent la reconnaissance des langues, la retranscription et la traduction à partir de sources audio. Ils sont susceptibles de fonctionner sur une cinquantaine des langues différentes mais les performances sont très dépendantes de celles-ci. Nous n'avons utilisé que la version la plus performante des modèles proposés. Dès la sortie de ce modèle, les chercheurs en sciences humaines se sont emparés de l'outil et les premiers tests ont à la fois montré son intérêt et quelques limites à son utilisation. Chitour (2023) explore par exemple l'utilisation de ce modèle pour la transcription d'entretiens, méthode utilisée régulièrement dans de nombreuses disciplines. Dans le même esprit, Wollin-Giering et al. (2024) comparent Whisper aux outils précédemment disponibles pour la transcription d'entretiens qualitatifs en allemand et en anglais. Ils concluent à la supériorité des résultats proposés par Whisper. Nous n'avons toutefois pas trouvé d'exemples d'utilisation sur le type de données qui nous intéresse ici.

Le tableau 2 présente les principales caractéristiques de corpus, du point de vue de la longueur des messages :

| | |
|---|---|
| Nombre de messages | 2210 |
| Durée moyenne | 33,3 secondes |
| Message le plus long | 19 minutes 41 secondes |
| Messages < 1 seconde | 5 |
| Messages > 60 secondes | 327 |

*Tableau 2 : description des messages du point de vue de leur durée*

Dans cette collection, le message le plus long s'étend sur 19 minutes. 327 messages durent plus d'une minute et seulement 5 sont inférieurs à 1 seconde. Ces messages très courts ne

[3] Nous avons utilisé la première version de ce modèle. Depuis nos analyses, deux autres versions ont été proposées, chacune améliorant les performances en comparaison de la version précédente.
[4] https://openai.com/research/whisper
[5] https://github.com/openai/whisper

contenaient aucune parole. La durée moyenne d'un message est de 33 secondes. Les messages laissés sur le fil « blabla » sont significativement plus longs (35,88 secondes en moyenne) que ceux du fil « InfoAction » (25,13 secondes en moyenne) ($U=360792$ ; $p < 0,0001$). Mis bout à bout, les messages du canal MVT31 représentent 20h28 d'enregistrement. Nous pouvons immédiatement souligner que la transcription manuelle de ces données aurait occupé un chercheur pendant plusieurs jours, voire plusieurs semaines. Sur une machine équipée d'une carte graphique avec 16Go de RAM, la transcription a durée environ 4h.

### *3.2. Première inspection des données*

Une première exploration des données produites permet de mettre en évidence quelques limites des transcriptions de Whisper. Deux exemples nous paraissent pouvoir compléter les premières remarques de Chitour (2023). D'une part la présence de transcription entièrement en anglais, alors que le locuteur s'exprime en français. Un des intervenants du canal a pris l'habitude de commencer ces messages oraux par l'expression « Good morning Toulouse ! » sur le ton utilisé par Robin Williams dans le film « Good Morning Vietnam » (Levinson, 1987). Le modèle a donc entièrement traduit le reste de son propos en anglais. Même si la traduction est de bonne qualité, ce n'est pas le résultat escompté. Nous avons éliminé les cinq longues interventions concernées dans certaines analyses. D'autre part, il nous semble que Whisper est victime d'hallucination, comme d'autres modèles du même genre. Rappelons que pour son fonctionnement, le modèle a été entraîné à reconnaître les sons des voix et à les distinguer des autres sources sonores potentielles. Mais que peut-il faire quand il n'y a pas de voix ou qu'elles sont presque inaudibles ? C'est ce qui s'est produit avec un enregistrement « fantôme ». Un utilisateur semble en effet avoir déclenché non-intentionnellement l'enregistrement d'un message sur *Telegram* avant de vaquer à ses occupations. Au début de l'enregistrement, une conversation est suffisamment proche du téléphone pour que la transcription soit fidèle aux paroles, mais les interlocuteurs s'éloignent rapidement et les voies se perdent dans le bruit de fond. Whisper a comblé ce manque par une longue série de «  per jansoi e per loro e per jansoi e per jansoi e per loro » représentant 526 occurrences. Ce message a été éliminé..

En dehors de ces deux exemples, nous retrouvons les commentaires de Chitour (2023) : Whisper élimine certaines incises, marques d'hésitation ou répétitions. Il omet certains adverbes, conjonctions ou prépositions et peut confondre des homophones. Il commet également des erreurs sur les noms propres. Ainsi, la piscine de Nakache devient la piscine d'Arkeche, la ville de Muret devient Muray et le quartier de Sesquières devient Cesquière. Les erreurs sur les noms propres sont probablement les plus fréquentes dans ce corpus qui contient beaucoup de références géographiques locales. Whisper en a même inventé : un message signalant le retard d'une personne nous indique qu'il « va être à Labour ». Nous avons par contre été surpris par la capacité du modèle à reconnaître les acronymes : OTAN, ONU, BFM TV, MEDEF, LCI et beaucoup d'autres sont parfaitement retranscrits (et en majuscules). Même si, comme le notait Chitour, le modèle peut être amené à faire des fautes de grammaire, la retranscription est très propre du point de vue de l'orthographe.

## 4. Comparaison entre écrit et oral

### *4.1. Statistiques lexicales*

Le tableau 3 permet de comparer, du point de vue des statistiques lexicales, le corpus des messages textuels et le corpus des retranscriptions des messages audios :

| | écrit | oral |
|---|---|---|
| nombre de textes | 50689 | 2195 |
| occurrences | 895914 | 220723 |
| nombre de formes | 22282 | 7222 |
| nombre d'hapax | 10239 | 3035 |
| % hapax (occurrences) | 1,14 % | 1,38 % |
| % hapax (formes) | 45,95 % | 42,02 % |
| taille moyenne d'un texte (en occurrences) | 17,67 | 100,56 |

*Tableau 3 : Statistiques des corpus écrit et audio après lemmatisation*

Le premier constat que nous pouvons faire porte sur la taille de ces messages. Les messages audios contiennent en moyenne 5 fois plus d'occurrences que les messages écrits (100,56 pour l'audio contre 17,7 pour l'écrit). Nous noterons que Fraisse et Breyton (1959) faisaient déjà cette remarque à propos de production d'enfants au milieu du siècle dernier.

À l'écrit, la moyenne d'occurrences par texte (17,7) s'approche de celle des tweets du temps où ils étaient limités à 140 caractères. Seul 1 % des messages présentent plus de 100 occurrences.

### *4.2. Comparaison des corpus à partir des spécificités*

Lorsqu'il compare des corpus oraux et écrits de plusieurs langues, Dewaele (2001) note la permanence de différences sur l'usage quantitatif de certains lexiques. Plus précisément, il montre que l'oral utilise significativement plus les pronoms, les conjonctions, les adverbes et les interjections quand l'écrit privilégie les substantifs, les adjectifs, les prépositions et les articles.

Nous avons soumis les catégories grammaticales de ces corpus à un calcul de spécificités. Dans IRaMuTeQ, c'est la bibliothèque lexicometry (TXM Team, 2013 ; Heiden, 2010) de R qui est utilisée pour ce calcul.

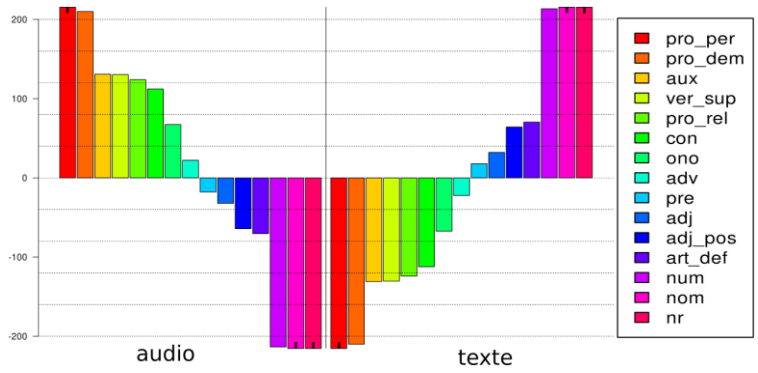


*Figure 1 : scores de spécificités des catégories grammaticales sur le tableau lexical entier partitionné selon le type de texte (Freq. Min. = 1). Les barres noires signalent des scores qui tendent vers l'infini.*

Nous retrouvons quasiment à l'identique les résultats de Dewaele (2001), même si nous devons les relativiser en prenant en compte le manque de précision des dictionnaires d'IRaMuTeQ et les particularités de la transcription automatique.

Du côté de l'oral, le score de spécificité des pronoms personnels est tiré par les formes *il*, *je*, *moi*, *on*, *vous* et *ils*. La spécificité des pronoms démonstratifs peut s'expliquer quasi intégralement par la surreprésentation du *c* et du *ça* avec une cédille alors que l'écrit est marqué par l'usage du *ca* sans cédille, commun à tous les corpus issus du numérique et qui pèsera sur les formes non reconnues (nr). Nous noterons également, toujours pour l'oral, le sur emploi des auxiliaires *être* et *avoir* auxquels nous pouvons ajouter certains des verbes marqués comme supplémentaires dans IRaMuTeQ (*dire*, *faire*, *savoir*, *vouloir*). Les pronoms relatifs sont portés par le triptyque *qui*, *que*, *quoi*, les conjonctions par *parce_que*, *donc* et *qu* et les adverbes par *vraiment* et *franchement*. Notons enfin que nous retrouvons surreprésentés à l'oral les marqueurs classiquement présents dans les entretiens retranscrits : le verbe *aller* et les formes *truc* et *chose*.

Les messages écrits présentent assez logiquement une surreprésentation des formes non reconnues. Il va s'agir d'une part, de toutes les abréviations utilisées dans ce mode de communication (*gj*, *lol*, *mdr*, *svp* etc...) et d'autre part, des mots mal orthographiés (comme le *ca*, mais également de tous les mots écrits sans leur accent : *meme*, *peage*, etc...). La surreprésentation des noms est portée par des formes comme *action*, *grève* et *péage*, mais l'effet le plus fort est généré par *rond* et *point*. C'est ici une conséquence de la qualité de la transcription. Dans celle-ci, Whisper met le tiret entre *rond* et *point*, la forme faisant partie du dictionnaire des expressions d'IRaMuTeQ, elle sera codée *rond_point*. A l'écrit, ce tiret est très majoritairement absent et *rond* et *point* pèseront le double de *rond-point*. La surreprésentation des chiffres et des nombres (num) est également potentiellement une conséquence de la transcription : Whisper code les informations chiffrées en toute lettre et elles seront alors classées dans les adjectifs numériques.

L'étape suivante de notre analyse va consister à réaliser une classification avec la méthode Reinert sur les données de ces deux corpus pour comparer les thématiques qu'ils abordent.

### *4.3. Classification sur les données écrites*

Étant donné la taille moyenne des productions écrites, nous n'avons pas segmenté ce corpus avant de le soumettre à une classification avec la méthode Reinert (Reinert, 1983 ; Ratinaud, 2018) :


*Figure 2 : Dendrogramme, taille des classes et profils des classes issues de l'analyse Reinert sur le corpus écrit*

Nous décrirons succinctement ces thématiques en suivant le dendrogramme de gauche à doite. Les classes 4 et 10 témoignent de l'organisation en temps réel de l'action d'occupation des ronds-points, des blocages des accès aux centres commerciaux ou aux raffineries et des opérations de péages gratuits. Nous retrouvons dans le lexique de ces classes des marqueurs géographiques de la ville de Toulouse et de ses environs (Capitole fait référence à la place de la mairie de Toulouse; Wilson est une autre place du centre-ville; Cyprien fait référence au quartier Saint-Cyprien; Socamil est une société de transport routier; Muret est une ville à une trentaine de kilomètres de Toulouse). L'originalité de ces lieux périphériques de mobilisation (en opposition à l'occupation classique des centres-villes) a été soulignée par Blavier et Walker (2022).

La classe 2 traite de la place de l'extrême droite dans le mouvement. On retrouve également tout un lexique lié à l'immigration, l'antisémitisme et d'habitude coutumier d'un positionnement discursif politique propre aux partis extrêmes. Contrairement à ce que le mouvement affiche en brandissant l'absence d'affiliation politique, nous relevons ici de fortes polarisations dans les discours qui témoignent d'une diversité d'ancrages politiques proches des extrêmes et non solubles dans les revendications qui leur sont communes.

Cette classe 2 est sur représentée à partir de début février. Lors de la manifestation du 2 février, un journaliste d'extrême droite, Vincent Lapierre, a prétendu avoir été attaqué à Toulouse, par des "black blocs" habituellement associés au mouvement d'extrême gauche et anarchiste. Cette attaque qui a par ailleurs été démentie par l'observatoire des pratiques policières, va donner lieu à des discussions sur Telegram, ce qui rendra saillante l'appartenance politique de l'administrateur et de certains membres de cette communauté qui vont prendre parti ou non pour Vincent Lapierre.

La classe 3 recoupe plusieurs sujets liés à l'actualité locale (le report du carnaval de la ville entre autres). La classe 6 souligne la dimension routinière des manifestations et de l'organisation logistique. Nous y retrouvons le lexique classique des mobilisations, de leur temporalité et de leur localisation (manif, manifester, samedi, paris…). La classe 7 témoigne de discussions d'auto défense vis-à-vis des violences policières, et aussi de conseils en matière de couverture de ces violences, notamment le fait de filmer les manifestations et les arrestations. Dans cette classe, on retrouve à la fois de l'anxiété et de l'agacement devant la répression policière, d'autant plus qu'elle a été minorée par les médias (Ratinaud & Sebbah, 2022). Rappelons que le niveau des violences policières à ce moment en France a conduit Michelle Bachelet, haut-commissaire aux droits de l'homme de l'ONU, à demander en urgence au conseil des droits de l'homme à Genève « une enquête approfondie sur tous les cas rapportés d'usage excessif de la force. » (Le Monde, 2019).

La classe 8 traite du pouvoir d'achat (payer, taxe, impôt salaire…) et s'articule à la classe 9 qui discute du mouvement, de son mode de fonctionnement, de sa structuration progressive (assemblée, vote, "ag" pour assemblée générale) et des revendications, dont le RIC (référendum d'initiative citoyenne). Nous retrouvons dans cette classe un lexique assez proche du registre des mouvements syndicaux. C'est aussi une classe qui traite de l'approche des élections européennes du 26 mai 2019. Toujours en lien, la classe 5 traite spécifiquement de la position des syndicats par rapport au mouvement. Enfin la classe 1 témoigne simplement des vocables mobilisés sur les espaces des réseaux socionumériques. C'est aussi la mise en perspective de Telegram par rapport aux autres réseaux sociaux.

### *4.4. Classification sur les données audios*

Pour cette analyse, le corpus oral a été découpé en 6881 segments de texte. La figure 4 présente les résultats de la classification :

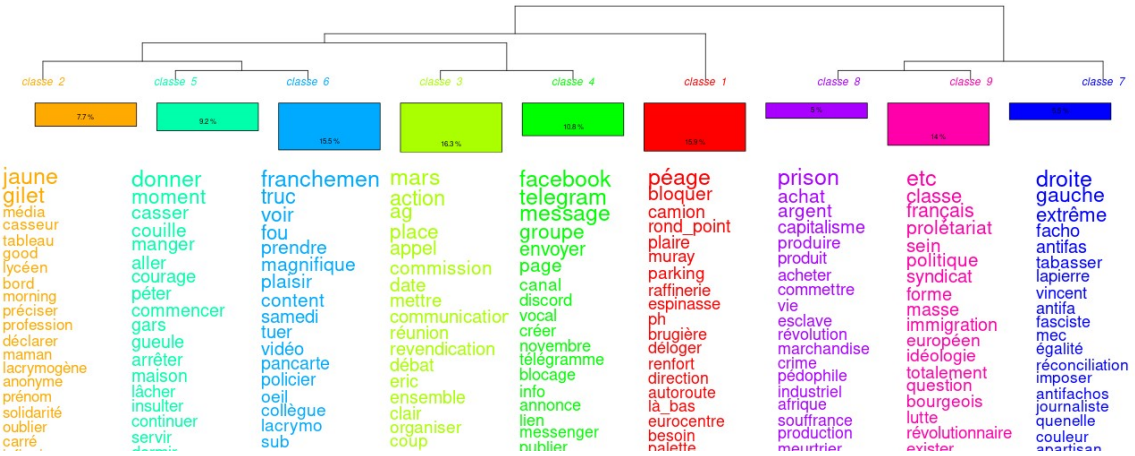


*Figure 4 : Dendrogramme, taille des tailles et profil des classes de l'analyse Reinert sur le corpus audio*

Une première lecture de ces résultats nous amène en territoire connu. Il nous semble effectivement identifier des thèmes communs entre le corpus oral et le corpus écrit. Pour objectiver cette impression, nous avons reconstruit un corpus avec l'ensemble des classes produites (écrit et oral), pour étudier les distances de Labbé (Labbé et Monière, 2000) entre elles.

### *4.5. Comparaison des classifications*

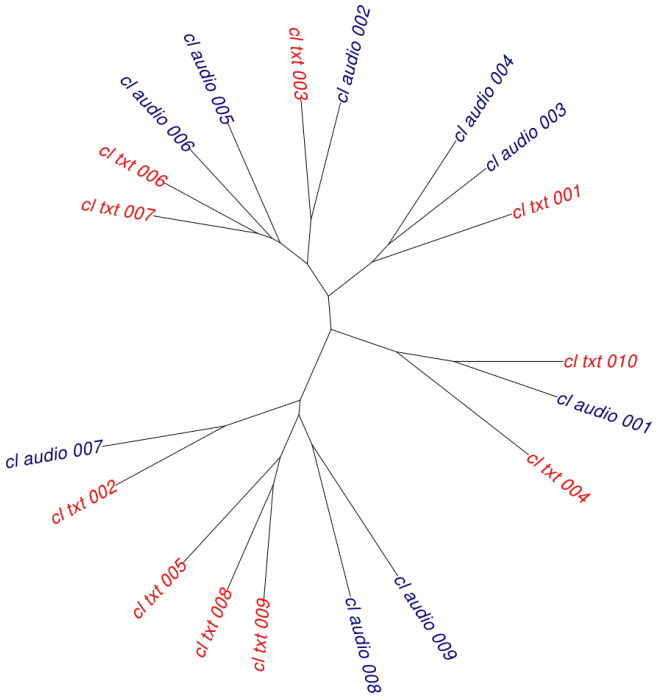


*Figure 5 : Arbre de la classification (méthode de Ward) sur la matrice des distances de Labbé entre les classes lexicales des deux analyses ; seules les formes actives participent (Ratinaud et Marchand, 2016)*

Sur la droite de la figure 5, nous voyons que le lexique du répertoire d'action des classes 4 et 10 de l'analyse textuelle se rapproche de la classe équivalente dans l'analyse des messages audios. Les fautes commises sur les noms de lieu (*Muray* pour *Muret*, *Espinasse* pour *Lespinasse* par exemple) sont gommées par les contextes similaires de leur apparition. Juste au-dessus, la classe 1 de l'écrit qui contient le lexique des réseaux socionumériques et des conseils sur la communication rejoint les classes 3 et 4 de l'audio. Les classes de témoignages sur les pratiques policières et les partages d'expériences des classes 6 et 7 des écrits se positionnent proches des classes équivalentes de l'audio (classes 5 et 6). En bas à gauche de la figure, les classes 7 de l'audio et 2 de l'écrit, qui traitent de la place de l'extrême droite dans le mouvement, apparaissent proches. Enfin, bien que positionnées dans le même espace, les classes traitant de la politique et du syndicalisme dans le corpus écrit (classes 5, 8 et 9) ne se confondent pas complètement avec les classes au lexique proche de l'audio. Il nous semble que des formes plutôt spécifiques du lexique de l'extrême droite et de l'extrême gauche sont plus présentes à l'oral qu'à l'écrit. Cette impression est confirmée par l'étude des spécificités lexicales : les formes *antifas*, *gauche*, *droite*, *prolétariat*, *musulman*, *extrême*, *capitalisme*, *fascisme*, *révolutionnaire*, *insoumis*, *drapeau*, *antifaschos*, *bourgeoisie*, *syndicat*... sont toutes surreprésentées à l'oral. Non seulement des thématiques équivalentes ou très proches sont identifiables entre les corpus oraux et écrits, mais nous noterons par ailleurs qu'elles apparaissent dans des proportions très proches.

## 5. Conclusion

Cela fait très longtemps que les chercheurs en sciences humaines attendent un outil fiable, économiquement viable et suffisamment respectueux de la vie privé pour être utilisé sur tous les types des corpus. Après cette première expérience, il nous semble que Whisper est tout à fait apte à remplir cette fonction. Contrairement à d'autres outils, nous n'avons pas eu à transférer nos données sur des serveurs extérieurs. La transcription que nous obtenons, même si elle n'est pas parfaite, et d'une qualité suffisante pour le type d'analyse que nous menons. Cela ouvre des perspectives, notamment dans l'analyse des productions sur les réseaux numériques, marquées par un nombre important de vidéos, mais également et plus généralement, pour l'analyse des médias, parce la télévision et la radio vont pouvoir rejoindre les analyses quantitatives produites sur la presse par exemple.

Le corpus sur lequel nous travaillons est d'une grande richesse pour l'étude des variations entre oral et écrit. Rares sont les corpus provenant de données spontanément produites, disponibles à la fois sous forme d'écrit et d'oral, issus de populations équivalentes et portant sur une thématique restreinte et une temporalité réduite. Malheureusement, la nature de ces données rend inenvisageable de les partager à grande échelle.

Enfin, sur le plan des statistiques textuelles, nous pouvons nous étonner des résultats du croisement des classifications. Corpus oraux et corpus textuels ont toujours été considérés comme des genres textuels différents et l'étude des spécificités entre ces deux types de texte tendrait à renforcer cette conviction. Mais beaucoup des différences observées peuvent être contrôlées dans une phase de nettoyage et de formatage des corpus, notamment les variations morphosyntaxiques, ce qui conduirait à encore plus de proximité lexicale entre les deux. Cette proximité est probablement surtout la conséquence du fait que, sur ce type de support, les productions écrites sont, dans leur structuration, très proches des productions orales.

## Bibliographie

Baisnée O., Cave A., Gousset C., Nollet J. et Parent, F. (2022). The digital coverage of the yellow vest movement as protest activity. *French Politics*, 20 (3-4), 529-549.

Chitour Y. (2023). *Whisper pour retranscrire des entretiens*, https://www.css.cnrs.fr/fr/whisper-pour-retranscrire-des-entretiens/

Cointet J.-P., Morales P. R., Cardon D., Froio C., Mogoutov A., Ooghe B. et Plique G. (2021). What colours are the yellow vests ? An ideological scaling of Facebook groups. *Statistique et Société*, 9 (1- 2), 79-107.

Dewaele J.-M. (2001). Une distinction mesurable : corpus oraux et écrits sur le continuum de la deixis. *Journal of French Language Studies*, 11 (2), 179–199. doi:10.1017/S0959269501000229

Fraisse P. et Breyton M. (1959). Comparaisons entre les langages oral et érit. *L'année psychologique*, 59 (1), 61-71

Heiden S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In K. I. Ryo Otoguro (Ed.), *24th Pacific Asia Conference on Language, Information and Computation – PACLIC24,* Institute for Digital Enhancement of Cognitive Development, Waseda University, Sendai, Japan, 389-398.

Labbé D. et Monière D. (2000). La connexion intertextuelle. Application au discours gouvernemental québécois. In M. Rajman et J.-C. Chappelier (Eds), *Actes des 5èmes Journées Internationales d'Analyse statistique des Données Textuelles*, Lausanne : EPLF, 85-94.

Levinson B. (1987). *Good Morning Vietnam*, Touchstone Pictures.

Marchand P., Sebbah B., Renard J., Cabanac G., Thiong-Kay L., Souillard N. et Loubère L. (2019). *Vrai débat : sortir du débat pour négocier*. (Rapport de recherche, LERASS, Université Paul Sabatier).

Radford A., Kim J. W., Xu T., Brockman G., McLeavey C. et Sutskever I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv*, https://arxiv.org/abs/2212.04356

Ratinau P. (2018). Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ. In D. F. Iezzi L. Celardo et M. Misuraca (Eds.), *JADT 2018, Proceedings of the 14th international conference on statistical analysis of textual data (Vol. 2)*, Rome, 616–625.

Ratinaud P. et Marchand P. (2016). Quelques méthodes pour l'étude des relations entre classifications lexicales de corpus hétérogènes : application aux débats à l'assemblée nationale et aux sites web de partis politiques. In D. Mayaffre, C. Poudat, L. Vanni, V. Magri et P. Follette (Eds), *Statistical Analysis of Textual Data,* 193–202.

Ratinaud P. et Sebbah B. (2022). Frénésie médiatique dans la presse quotidienne française : analyse lexicométrique de plus de 100 000 articles sur les Gilets jaunes. In A. Mercier et J.-M. Charon (Eds.), *Les Gilets jaunes : un défi journalistique,* Paris : Panthéon-Assas, 93–113.

Reinert M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les Cahiers de l'Analyse des Données*, 8 (2), 187-198.

Sebbah B., Marchand P., Loubère L., Souillard N., Smyrnaios N. et Renard J. (2019). Les gilets jaunes : le pari gagné de l'existence médiatique ? (Rapport de recherche, LERASS, Université Paul Sabatier).

Sebbah B. et Ratinaud P. (2023). *The Organization of a Movement on Telegram: The Yellow Vests in France*. 2023 American Political Science Association Annual Meeting & Exhibition, Los Angeles.

Souillard N., Sebbah B., Loubère L., Thiong-Kay L. et Smyrnaios N. (2020). Les Gilets jaunes, étude d'un mouvement social au prisme de ses arènes médiatiques. *Terminal. Technologie de l'information, culture & société*, 127. DOI: 10.4000/terminal.5671

TXM Team (2013). *TXM Manual*. ICAR Laboratory, Lyon University & CNRS, Lyon, France.

Wollin-Giering S., Hoffmann M., Höfting J. et Ventzke C. (2024). Automatic Transcription of English and German Qualitative Interviews. *Forum : Qualitative Social Research / Qualitative Sozialforschung*, 25 (1), 335–371.