Les temps de la crise sanitaire au prisme d'une série chronologique : une étude phonético-textométrique

Sascha Diwersy, Ivana Didirková, Christelle Dodane, Fabrice Hirsch,

Les temps de la crise sanitaire au prisme d'une série chronologique : une étude phonético-textométrique

Sascha Diwersy[1], Ivana Didirková[2], Christelle Dodane[1], Fabrice Hirsch[1], Giancarlo Luxardo[1]

[1]Praxiling UMR 5267 – prenom.nom@univ-montp3.fr

[2]DEPA, Université Paris 8 – ivana.didirkova@univ-paris8.fr

Abstract

We present a corpus built upon the broadcast addresses of president Emmanuel Macron, delivered during the health crisis in 2020 and 2021. We integrate the orthographic transcriptions and the digitized audio recordings. This corpus is tokenized on the basis of both its text content and the related acoustic signal. The transcriptions then get a morphosyntactic tagging, a lemmatization and an annotation according to dependency relations. As a result, we make use of a linguistic environment organized according to a data model integrating acoustic, prosodic, (morpho)-syntactic and lexical features, thus allowing a processing both with textometric software, such as TXM, and with speech analysis software.

Keywords: textometry, written corpora, spoken corpora, exploratory data analysis.

Résumé

Nous avons construit un corpus constitué par les neuf allocutions télévisées du président Emmanuel Macron, prononcées pendant la crise sanitaire en 2020 et 2021, en intégrant les transcriptions orthographiques et les enregistrements audio numériques. Ce corpus est soumis à une segmentation textuelle mais aussi une segmentation à partir du signal acoustique. Successivement, les transcriptions ont fait l'objet d'un étiquetage morpho-syntaxique, d'une lemmatisation et d'une annotation au niveau des relations de dépendance. L'ensemble a été fusionné sous forme d'une base textuelle selon un modèle de données associant des traits acoustiques, prosodiques, (morpho-)syntaxiques et lexicaux, et permettant ainsi une exploitation aussi bien par des logiciels textométriques, comme TXM, que des outils d'analyse de la parole.

Mots clés : textométrie, corpus écrits, corpus oraux, analyse exploratoire des données

1. Introduction

L'analyse de discours politiques représente l'un des premiers domaines historiquement ciblés par la textométrie. Dans ce domaine, l'étude des évolutions chronologiques revêt un intérêt majeur. C'est ainsi qu'ont été forgées les notions de « série textuelle chronologique » pour caractériser des corpus textuels ordonnés selon le temps, respectant certains critères d'homogénéité, et de « temps lexical » pour décrire les variations (accroissements, disparitions ou discontinuités), observées dans le lexique (Salem, 1988). Afin d'étudier de tels corpus, les techniques d'analyse de données multivariée peuvent être utilisées : elles se fondent en principe sur une mesure de l'écart par rapport à une distribution régulière attendue. D'autres méthodes plus spécifiques ont été proposées, notamment des méthodes récursives permettant d'aboutir à une périodisation du corpus et à l'identification de champs lexicaux spécifiques.

Dans ce domaine des séries textuelles chronologiques, rares sont les tentatives de traiter des corpus intégrant des données sur la parole ou l'interaction. Pourtant, les textes politiques résultent souvent de transcriptions de l'oral, qu'il s'agisse d'interventions monologales ou d'interactions verbales entre différents interlocuteurs. C'est la raison pour laquelle l'objectif de cette proposition est de prendre en compte différents paramètres phonétiques et prosodiques (les pauses notamment) en vue de mettre en évidence un « temps phonétique », dont les continuités ou discontinuités seraient éventuellement synchrones avec le temps lexical.

2. Outils et traitements

Afin d'intégrer les deux sources, les transcriptions (données écrites) et les enregistrements audio (données orales), en un corpus unique, nous avons effectué les traitements suivants :

1. Collecte des versions vidéo et audio des allocutions de Macron à partir de la chaîne *YouTube* de France 24 par l'intermédiaire du navigateur *Firefox* avec l'extension *YouTube Video and Audio Downloader* [1] ;
2. Collecte de la version transcrite des mêmes allocutions, telles que diffusées sur le site de l'Elysée via la page https://www.elysee.fr/toutes-les-actualites ;
3. Traitement de la version audio au moyen du logiciel d'analyse de la parole *Praat* (Boersma et Weenink, 2021) pour la détection automatique d'unités interpausales et création d'une version TextGrid par l'alignement avec les transcriptions récoltées[2] ;
4. Conversion, par le logiciel *TEICONVERT* (Parisse *et al.*, 2020), de la version TextGrid des allocutions au format XML TEI_CORPO[3], conforme au module TEI de transcription de la parole ;
5. Enrichissement de la version XML par l'ajout d'annotations (morpho-)syntaxiques (parties du discours, traits morphologiques, relations de dépendances) [4] et de lemmes au moyen de l'analyseur *Stanza* (Qi *et al.*, 2020) ;
6. Projection des transcriptions phonétiques incluant une délimitation syllabique répertoriées dans le dictionnaire électronique *GLAWI*[5] (Hathout & Sajous, 2016) aux formes issues de la tokenisation par Stanza ;
7. Transformation des documents XML annotés au format VRT (verticalized text) servant d'entrée à l'indexation par CQP (Corpus Query Processor) (Evert, 2021) ;
8. Importation dans le logiciel de textométrie *TXM* dans le mode CQP.

Le corpus résultant, appelé *Covidis9* (Covid-19 : discours présidentiels)[6], est donc soumis à une annotation lexicale, syntaxique et phonétique et analysé selon les méthodes courantes de l'exploration textométrique, tout en y incluant des variables acoustiques telles que la durée des

[1] https://addons.mozilla.org/fr/firefox/addon/youtube_downloader_webx/

[2] Les versions écrites diffusées par l'Elysée présentant quelques divergences, les transcriptions utilisées pour l'alignement ont fait l'objet d'une adaptation aux paroles réellement énoncées dans l'allocution retransmise à la télévision.

[3] https://ct3.ortolang.fr/teiconvert/

[4] Suivant le schéma Universal Dependencies (UD ; documenté, pour le français, à la page https://universaldependencies.org/fr/index.html)

[5] Téléchargeable à l'adresse http://redac.univ-tlse2.fr/lexiques/glawi.html.

[6] Le corpus est consultable en libre accès sur la plateforme ORTOLANG à l'adresse : https://www.ortolang.fr/market/corpora/covidis9

pauses ou le débit de parole. Les variables acoustiques sont intégrées à la fois aux annotations des unités lexicales dérivant de la tokenisation des textes et aux propriétés de structure déduites de la structuration XML.

3. Structuration et encodage du corpus

3.1 Modèle conceptuel

Le modèle conceptuel utilisé pour notre étude dérive de celui couramment défini en textométrie. Celle-ci structure un corpus en textes composés d'unités lexicales et associe des propriétés aussi bien aux unités lexicales qu'aux structures de plus haut niveau telles que les phrases, les textes ou le corpus dans sa globalité.
Dans le cas de notre corpus d'étude, ce modèle se caractérise par l'ajout, à côté des unités de texte et de phrase, de deux unités de structure identifiées suite au traitement phonétique:

- la pause : un silence dans le signal acoustique,
- l'unité inter-pausale : une unité de la parole segmentée dans un énoncé et délimitée par des pauses

De par l'intégration des unités acoustiques en question, nous définissons deux propriétés de structures spécifiques : les propriétés colligationnelles[7] et les propriétés d'empan. Les propriétés colligationnelles permettent de décrire la position d'une unité de structure par rapport aux unités de structures de niveaux supérieurs dans la hiérarchie (dans l'ordre ascendant : pause ou unité inter-pausale – phrase – texte). Les propriétés d'empan caractérisent les unités de structure en termes de durée (mesurée en millisecondes), de vitesse d'articulation ou de débit (calculées en rapportant le nombre de syllabes des mots regroupés par une unité de structure à la durée de cette même unité) le cas échéant. Le tableau 1[8] donne les caractéristiques statistiques globales du corpus avec ses sous-échantillons.

3.2 Encodage des unités de structure et propriétés

Les concepts précédents sont encodés dans les documents résultants des différentes étapes de notre chaîne de traitement et, en dernier ressort, par le processus d'importation dans TXM. Dans l'ordre des traitements opérés, les unités de structure sont : (1) les allocutions (le texte) correspondant aux documents collectés, (2) les pauses et unités inter-pausales identifiées au moyen de Praat, (3) les phrases délimitées lors de la segmentation automatique des échantillons textuels opérée en amont de l'analyse syntaxique mise en œuvre par la chaîne d'annotation du logiciel Stanza.
Pour les allocutions, nous avons retenu, à côté d'un identifiant posé par défaut, la date de diffusion. Les jeux de descripteurs associés aux phrases, unités inter-pausales et pauses sont conçus de manière comparable et incluent les traits colligationnels ainsi que les propriétés de durée, de vitesse d'articulation (unités inter-pausales) ou de débit (phrases). Afin de pouvoir être traitées par TXM (en tant que variables catégorielles), la durée, la vitesse d'articulation et le débit ont été représentés aussi bien par leur valeur brute (variable numérique) que par l'une des trois propriétés associées aux trois terciles résultant d'une subdivision en trois intervalles

[7] Nous faisons appel à la notion de colligation textuelle, telle que proposée par Hoey (2005), qui décrit la préférence positionnelle d'une unité lexico-grammaticale dans un texte donné. Elle est ici généralisée à des unités caractérisées également par des traits phonétiques.

[8] Les tables et figures illustrant le présent article sont consultables à l'adresse suivante :
https://www.ortolang.fr/market/corpora/covidis9?path=%2FJADT2022
(à noter que dans la première version du corpus, utilisée ici, la neuvième allocution n'a pas été incluse).

égaux. Les bornes d'intervalle qui délimitent les groupes de valeurs pour les différentes unités avec les propriétés concernées sont précisées dans le tableau 2. Le tableau 3 donne un aperçu de l'encodage des unités de structure du corpus avec leurs propriétés et les modalités associées. Pour ce qui est des propriétés associées aux unités lexicales, celles-ci correspondent aux annotations obtenues au moyen de la chaîne de traitement Stanza (catégorie grammaticale, lemme, traits morphologiques et relation de dépendance syntaxique) avec, en extension, une réplication des propriétés de la tête syntaxique de chaque token et ses traits de colligation textuelle (au niveau du texte entier, des phrases et des unités inter-pause). En plus de leur modélisation comme unité de structure, les pauses ont également été associées à des propriétés au même niveau que les unités lexicales. Le modèle des propriétés lexicales du corpus est documenté par le tableau 4.

4. Résultats

Nous présentons d'abord deux traitements analytiques réalisés dans TXM sur notre corpus, l'un prenant en compte des propriétés d'unités lexicales, l'autre des propriétés d'empan (durée des pauses, débit au niveau des phrases).

Nous réalisons deux Analyses Factorielles des Correspondances (AFC) sur les lemmes du corpus présents avec une fréquence supérieure ou égale à 4 :

- dans la première AFC (figure 1), par filtration de certains mots grammaticaux (ponctuation, déterminants, prépositions, numéraux, auxiliaires, conjonctions, pronoms, outre les tokens catégorisés 'X' correspondant aux pauses) en appliquant la requête : [upos!="(PUNCT|DET|ADP.*|NUM|AUX|.CONJ|PRON|X)"]. L'étude des résultats conduit à distinguer un regroupement des allocutions en quatre périodes (deux sur l'axe 1, deux sur l'axe 2).
- à partir des 25 lemmes fournissant les plus fortes contributions sur l'axe 1, une deuxième AFC (figure 2) accentue l'opposition entre les premières allocutions (mars 2020) et celles de sorties de crise (mars 2021 et juillet 2021 – le point-colonne 2021-07-12 fournissant la meilleure contribution relative sur le plan (1,2)). Par rapport à un ordonnancement « attendu » des différentes allocutions, la représentation obtenue fait ressortir ici la permutation entre les deux premières de mars 2020 (celle du 2020-03-16 se distinguant de toutes les autres) et la position « aberrante » de l'allocution du 2020-06-14.

L'analyse de spécificités illustrée par la figure 3 permet de visualiser la distribution des durées des pauses dans les différentes allocutions (les durées sont classées en trois catégories C1, C2 et C3, pour durées courtes, moyennes ou longues). On remarquera que les pauses longues sont surreprésentées dans les trois premières allocutions et qu'elles deviennent rares à partir d'octobre 2020.

La distribution spécifique des pauses longues et courtes se reflète dans les différences très marquées du débit des phrases, qui font apparaître la même configuration mettant en opposition les allocutions de la première vague épidémique à celles à partir de la deuxième vague, avec une sur-représentation des phrases à débit ralenti (S_rate_C1) pour les unes et une sur-représentation des phrases à débit accélérée (S_rate_C3) pour les autres (figure 4).

D'après Duez (1999), les différences concernant la longueur des pauses ainsi que le débit et la vitesse d'articulation dans la parole politique peuvent être considérées comme étant des marques traduisant des situations au pouvoir différentes. Pauses brèves et débit accéléré indiquent la volonté d'optimiser l'exploitation du temps de parole pour accumuler arguments et idées exprimés – c'est le discours de la persuasion et de la (re)conquête du pouvoir. Pauses longues

et débit ralenti, en revanche, reflètent la position des dominants, et participent en dernière analyse de la construction de l'ethos présidentiel – c'est le discours du pouvoir qui s'exerce.

Dans le cas des allocutions de Macron, les différences observées quant aux traits acoustiques en question relèvent bien de préalables stratégiques divergents. Au début de la pandémie, face à ce qui allait devenir la première vague, Macron apparaît politiquement affaibli et en quête d'une légitimité largement perdue (cf. Sadoun-Kerber & Wahnich, 2022). Les premières allocutions vont être l'occasion de restaurer son autorité, restauration qui passe par la reconstruction de l'ethos présidentiel et l'activation de certains mythes constitutifs du mandat suprême de la Ve République, qui sont relayés par les thèmes marqués par les mots spécifiques de ces interventions télévisées en mars et avril 2020 : le mythe de l'homme providentiel, capable de passer à l'initiative face à la crise [2022-03-12 : *prendre* (indice de spécificité : 3,5) et *mesure* (2,6), associés dans la locution verbale *prendre des mesures*, *réagir* (2,7)], mythe central modulé par celui du "Père" (Sadoun-Kerber & Wahnich 2022), garant du bien-être des concitoyens [2022-03-12 : *santé* (2,8)] ; le mythe du "Chef" (Sadoun-Kerber & Wahnich 2022) légitimé à responsabiliser ses concitoyens [2022-03-12 : *compter* (4,7) dans l'emploi avec COI marqué par *sur*] et à orchestrer le contrôle de leurs comportements [2022-03-16 : *consigne* (2,9) ; 2022-04-13 : *règle* (2,6)], voire celui du "chef des armées" (ibid.) [2020-03-12 : *mobilisation* (2,1) ; 2020-03-16 : *guerre* (6,2)]. A partir de la deuxième vague pandémique, d'abord sous-estimée pour son ampleur, Macron est confronté au risque de perdre le crédit supposément acquis grâce à sa gestion de crise, et il se voit obligé de persuader la population du bien-fondé de l'action gouvernementale. Notons entre autres que parmi les termes spécifiques de l'allocution du 28 octobre 2020, on trouve désormais à côte du nom *mesure* (2,9), qui renvoie en partie au confinement mis en place pendant la première vague [*printemps* (5,8)], celui de *stratégie* (2,7), qui souligne le caractère rationnel de son action. Cette même rationalité se voit d'ailleurs déclinée à travers un ensemble de séquences descriptives et argumentatives, cadrées de manière méta-discursive par une série de phrases interrogatives partielles marquées par le terme interrogatif *quel* (2,8).

Pour confirmer que les oppositions entre allocutions que nous venons d'observer aussi bien sur le plan acoustique que sur le plan lexical ne relèvent pas d'un parallélisme aléatoire, il suffit de mettre en œuvre des requêtes qui ciblent des unités combinant les deux niveaux de traits. Si l'on interroge par exemple, au sein d'une partition par allocution, la distribution des mots en conjonction avec le débit de la phrase dans laquelle ils sont employés, on retrouve non seulement le même regroupement de textes (cf. l'AFC en figure 5) qu'avec les calculs déployés séparément, mais aussi en grande partie les mêmes items spécifiques.

Notons, pour terminer ce bref aperçu de résultats textométriques, que les différences de spécificités acoustiques observées se reflètent également de par la répartition au fil des allocutions des traits concernés, en l'occurrence le débit des phrases. Ainsi, on constate que le contraste du profil acoustique des différents groupes d'allocutions identifiés supra se confirme à travers les colligations textuelles des phrases à débit ralenti et accéléré. Les courbes de progression obtenues pour les allocutions du 2020-03-16 et du 2020-10-28 (figure 6), qui se distinguent par les distributions les plus opposées, montrent une répartition nettement différente des deux catégories de phrases selon les parties initiale, médiane et finale de chacune de ces allocutions, avec des pics locaux à des endroits bien différents. A côté d'une « dominante » acoustique, établie en termes de sur-emploi respectif des phrases à débit ralenti (76 occurrences, score de spécificité de 6 pour le 2020-03-16 – voir figure 4) ou accéléré (110 occurrences, score de spécificité de 14 pour le 2020-10-28), on voit donc apparaître, à l'égard des allocutions concernées, un cadençage caractéristique (une distribution plus régulière pour le débit « dominant »), dont l'analyse détaillée sera réservée à une future contribution.

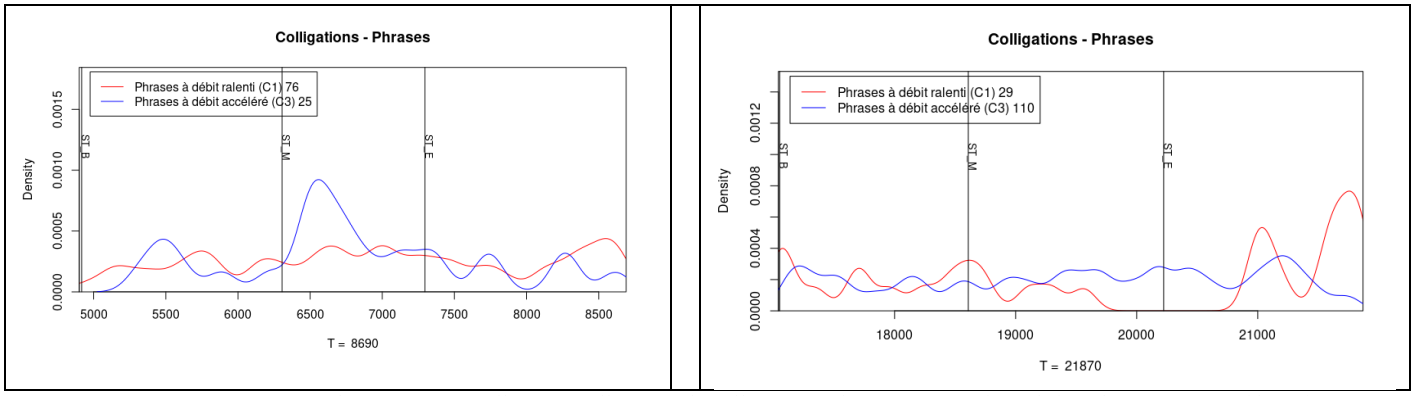

Figure 6 - Progression des propriétés colligationnelles pour les allocutions du 2020-03-16 (gauche) et du 2020-10-18 (droite)

Conclusion

Avec cette étude, nous proposons donc de démontrer que les méthodes de l'analyse exploratoire des données mises en œuvre par le logiciel TXM peuvent s'appliquer non seulement à des unités lexicales et leurs propriétés morphosyntaxiques mais aussi à des traits phonétiques tels que les pauses, la vitesse d'articulation ou le débit de la parole et que la proximité des classifications qui résultent des opérations appliquées fournissent des indices utiles dans l'interprétation. Parmi les extensions que nous envisageons à présent pour notre modèle, nous pouvons mentionner l'intégration de caractéristiques prosodiques.

Références

Boersma, Paul & Weenink, David (2021). *Praat: doing phonetics by computer* [Computer program]. Version 6.2.04, retrieved 18 December 2021 from http://www.praat.org/

Duez, Danielle (1999). La fonction symbolique des pauses dans la parole de l'homme politique. *Faits de langues* 7(13), 91–97. https://doi.org/10.3406/flang.1999.1242.

Evert, Stefan & The CWB Development Team (2021). *The IMS Open Corpus Workbench (CWB) - CQP Interface and Query Language Manual*, retrieved 10 March 2022 from https://cwb.sourceforge.io/files/CQP_Tutorial.pdf.

Hathout, Nabil & Sajous, Franck (2016). Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary. *Proceedings of the tenth International Conference on Language Resources and Evaluation* (LREC 2016), pp. 136-1376, Portorož, Slovenia.

Hoey, Michael (2005). *Lexical priming: a new theory of words and language*. London ; New York: Routledge.

Parisse, Christophe, Etienne, Carole & Liégeois, Loïc. (2020). TEICORPO: a conversion tool for spoken language transcription with a pivot file in TEI, Journal of the Text Encoding Initiative. TEI Consortium. https://halshs.archives-ouvertes.fr/halshs-03043572.

Qi, Peng, Zhang, Yuhao, Zhang, Yuhui, Bolton, Jason & Manning, Christopher D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

Sadoun-Kerber, Keren & Wahnich, Stéphane (2022). Emmanuel Macron face à la Covid-19 : un Président en quête de réparation d'image. *Argumentation et analyse du discours* 28. https://doi.org/10.4000/aad.6113. http://journals.openedition.org/aad/6113 (11 May, 2022).

Salem, André (1988). Approches du temps lexical, *Mots* 17, 105–143. https://doi.org/10.3406/mots.1988.1401.