

Université de Neuchâtel
Centre de linguistique de Corpus

Corpus
Les discours des présidents de la République française
(1958-2018)

Avril 2019

dominique.labbe@umrpacte.fr

cyril.labbe@imag.fr

denis.moniere@umontreal.ca

Ensemble des interventions publiques disponibles des présidents de la République française de mai 1958 à décembre 2019 : allocutions, entretiens, conférences de presse, messages...

En avril 2019, il comporte 8 748 textes, soit 18,431 millions de mots et un vocabulaire 45 575 vocables différents :

- les interventions radio-télévisées : textes effectivement prononcés par les présidents ;
- les autres discours et messages : pour la plupart, il s'agit de textes communiqués par les services de la présidence et diffusés par les agences de presse. Quelques textes sont des transcriptions d'enregistrements audio ou audiovisuels.

Ces documents ont été produits dans le cadre des fonctions de président, diffusés par la présidence et largement publiés dans la presse. Ils sont donc disponibles sans restrictions pour une utilisation non commerciale.

Contenu du dossier

Le tableau ci-dessous synthétise, pour les 7 présidents élus depuis 1958, le nombre d'interventions présents dans la base, le nombre total de mots (N) et le vocabulaire (V).

	Dates	Textes	N	V
Charles de Gaulle**	1958-1969	459	410 492	9 002
Georges Pompidou**	1969-1974	137	259 918	8 076
Valéry Giscard d'Estaing**	1974-1981	191	660 560	9 535
François Mitterrand	1981-1995	2 547	5 576 244	25 695
Jacques Chirac	1985-2007	2 478	4 081 697	24 054
Nicolas Sarkozy	2007-2012	1 074	3 221 250	21 597
François Hollande	2012-2017	1 544	3 182 939	20 403
Emmanuel Macron	2017-2018	318	1 037 988	14 819
Total		8 748	18 431 088	45 575

** Ces corpus seront complétés grâce aux archives des agences de presse et de la présidence (déposées aux Archives nationales).

Pour chaque président, deux sous-corpus sont distingués, grâce à des listes placées en tête du dossier :

Corpus A. Ensemble des interventions radiotélévisées (allocutions, entretiens, conférences de presse) ;

Corpus B. Autres interventions (allocutions, entretiens notamment avec la presse écrite, conférences de presse, messages...)

Ils sont décrits dans "0 Corpus....xls", placé en tête du dossier et lisible avec n'importe quel tableur.

Les textes sont classés par ordre chronologique (année, mois, jour). Exemple : DeGaulle19580515 est un communiqué émis le 15 mai 1958.

Les trois premières lignes de chaque fichier donnent :

- les indications bibliographiques (auteur, titre, lieu et date)
- le site où a été téléchargé le document électronique ou la référence du document papier auquel on peut se reporter,
- le nom de la personne ayant réalisé les opérations de saisie, de standardisation orthographique, de balisage et de lemmatisation avec la date de ces opérations.

Chaque texte est sous deux formats :

- le texte original (NomdeFichier.txt). La graphie de chaque mot a été contrôlée, notamment celle des noms propres. Tout le paratexte est placé entre <> (références du texte, questions des journalistes...).

- une version lemmatisée (même nom de fichier précédé du radical CN).

Dans le fichier CN, chaque mot du texte apparaît sur une ligne avec sa graphie standard, son entrée de dictionnaire (par exemple, l'infinitif du verbe) et son code grammatical, chaque champ est séparé par une virgule. La liste des codes grammaticaux est donnée dans le fichier "Nomenclature CN.pdf".

Chaque texte est accompagné d'un index (même nom de fichier précédé du radical II). Cet index est utilisé par les différents outils informatiques d'analyse des corpus.

Le vocabulaire du corpus est donné dans :

- index alphabétique (nom de corpus précédé du radical IA), lisible sous excel. Dans les deux premières colonnes, les mots du corpus - transcrits sous leurs graphies standards - sont rangés sous leurs entrées de dictionnaire (par ordre alphabétique). En troisième colonne, l'effectif dans l'ensemble du corpus. Dans les colonnes suivantes, les textes où ils apparaissent avec leurs nombres d'occurrences (effectifs absolus).

- index hiérarchique (nom de corpus précédé du radical IH) : principaux renseignements sur le vocabulaire du corpus avec les vocables les plus utilisés rangés par catégories grammaticales.

Règles d'usage :

Les corpus étiquetés sont couverts par la licence BSD
(<http://www.opensource.org/licenses/bsd-license.html>), Copyright (c) 2017, Cyril Labbé,
Dominique Labbé, Denis Monière.

Ces fichiers ne peuvent faire l'objet d'aucun usage commercial.

Les premières lignes ne doivent pas être enlevées.

Toute publication réalisée avec ces fichiers indiquera la source et sera envoyée à Dominique Labbé (adresse électronique au début de ce document).

Les outils d'exploitation des fichiers CN sont disponibles auprès de Dominique Labbé.

Si vous détectez des erreurs dans la saisie ou l'étiquetage, merci de prévenir Dominique Labbé afin qu'elles soient corrigées.

Bibliographie

(Principales études réalisées avec ces corpus)

La plupart des documents référencés ci-dessous sont disponibles gratuitement sur researchgate et les archives ouvertes (HAL) du CNRS.

- Les normes de standardisation et de lemmatisation utilisées pour produire le corpus

Dominique Labbé. *Normes de saisie et de dépouillement des textes politiques. Cahier du CERAT n° 7*. Grenoble : CERAT-IEP, avril 1990.

• Ouvrages

Edward Arnold, Cyril Labbé, Dominique Labbé & Denis Monière. *Parler pour gouverner : Trois études sur le discours présidentiel français*. Grenoble : Laboratoire d'Informatique de Grenoble, 2016, p. 38-53.

Dominique Labbé. *Le vocabulaire de François Mitterrand*. Paris : Presses de la Fondation Nationale des Sciences Politiques, 1990.

• Rapports, communications et articles (ordre chronologique inverse)

Dominique Labbé. Soixante-ans de discours présidentiels français (1958 – 2018). Qu'est-ce qui singularise Emmanuel Macron ? *Séminaire Mathématiques et société*. Université de Neuchâtel, 17 mai 2019.

Denis Monière & Dominique Labbé. Le vocabulaire des campagnes électorales. In Iezzi Domenica F., Celardo Livia, Misuraca Michelangelo. *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*. Roma: UniversItalia, 2018, p. 522-540.

Cyril Labbé & Dominique Labbé. *La répartition du vocabulaire*. Grenoble : Laboratoire d'informatique de Grenoble, septembre 2017.

Cyril Labbé & Dominique Labbé. Le chiffre dans le discours politique français contemporain. V. Giscard d'Estaing et les autres présidents. In Banks David. *La quantification dans le texte de spécialité*. Paris : L'Harmattan, 2016, p. 53-75.

Cyril Labbé & Dominique Labbé. Existe-t-il un langage propre à la politique ? In David Banks (Ed). *Aspects linguistiques du texte politique*. Paris : L'Harmattan, 2014, p. 7-28.

- Cyril Labbé & Dominique Labbé. La modalité verbale en français contemporain. Les hommes politiques et les autres. In David Banks (Ed). La modalité, le mode et le texte spécialisé. Paris : L'Harmattan, 2013, p. 33-61.
- Cyril Labbé & Dominique Labbé. La diachronie dans le discours politique. Le général de Gaulle. In David Banks (Ed). *Aspects diachroniques du texte de spécialité*, Paris, l'Harmattan, 2010, p. 129-148.
- Dominique Labbé. Le général de Gaulle en campagne. In David Banks (Ed.). *Aspects linguistiques du texte de propagande*. Paris : L'Harmattan, 2005, p. 213-233.
- Pierre Hubert, Cyril Labbé & Dominique Labbé. Automatic Segmentation of Texts and Corpora. *Journal of Quantitative Linguistics*, december 2004, 11-3, p. 193-213.
- Dominique Labbé. La richesse du vocabulaire politique : de Gaulle et Mitterrand. In Sylvie Mellet & Marcel Vuillaume. *Mots chiffrés et déchiffrés. Mélanges offerts à Etienne Brunet*. Paris : Champion, 1998, p. 173-186.
- Dominique Labbé. La France chez de Gaulle et Mitterrand. In Pierre Fiala et Pierre Lafon (dir). *Des mots en liberté. Mélanges Maurice Tournier*. Fontenay-aux-Roses : ENS Editions, 1998, p. 183-193.
- Dominique Labbé. Le "nous" du général de Gaulle. *Quaderni di studi linguistici*. 4/5, 1998, p 331-354.
- Fabre Cécile, Habert Benoît & Dominique Labbé.. La polysémie dans la langue générale et les discours spécialisés. *Sémiotiques*. 13, décembre 1997, p. 15-30.
- Pierre Hubert & Dominique Labbé. La structure du vocabulaire du général de Gaulle. In Sergio Bolasco, Ludovic Lebart et André Salem. *III Giornate internazionali di Analisi Statistica dei Dati Testuali*. Rome : Centro d'Informazione e stampa Universitaria, 1995, tome II, p. 165-176.
- Dominique Labbé. Les métaphores du général de Gaulle. *Mots*. 43, juin 1995.
- Pierre Hubert & Dominique Labbé. Vocabulary Richness. *Communication au Colloque de l'ALLC-ACH*. Paris, 19-23 avril 1994.
- Pierre Hubert & Dominique Labbé. La répartition des mots dans le vocabulaire présidentiel. *Mots*, n° 22, mars 1990, p. 80-88.
- Dominique Labbé. Des réformes à la cohabitation. Les quatre périodes du premier septennat Mitterrand. *Mots*. n° 22, mars 1990, p. 62-78.
- Pierre Hubert & Dominique Labbé. Un modèle de partition du vocabulaire. In LABBE Dominique, SERANT Daniel et THOIRON Philippe. *Etudes sur la richesse et la structure lexicales*. Genève-Paris : Slatkine-Champion, avril 1988, p. 93-114.