Présentation des corpus

Linguistique de corpus : Introduction

Chloé Tahar

29 septembre 2025

Les corpus

1. Corpus Wikipédia
2. Corpus Élysée
3. Corpus ESLO
4. Corpus BFM

Corpus Wikipédia

Corpus Wikipédia

WIKIPÉDIA
L'encyclopédie libre

Un corpus réalisé par l'ISLa de l'Université de Neuchâtel d'environ 10 millions de mots, qui récupère les 200 premières pages populaires des principaux projets (Arts & Culture, Biologie, Géographie, Histoire, Politique, Religion, Sciences, Société, Sport, Technologie) du Wikipédia français.

Variation par domaine de connaissance

Projet	N mots
Arts	950 759
Biologie	720 980
Géographie	864 021
Histoire	1 388 046
Politique	1 219 535
Religions	955 924
Sciences	771 712
Société	1 179 422
Sport	1 049 911
Technologies	813 337

Encyclopédie ou média d'information?

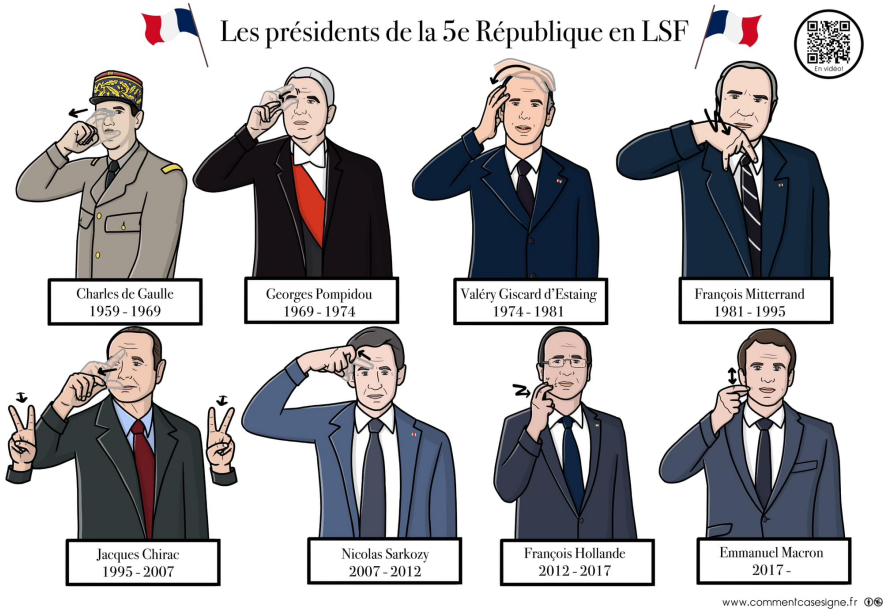
Doutreix (2010)[a]

- Wikipédia est une forme nouvelle, à la frontière du genre médiatique et du genre encyclopédique
- Le principe de neutralité de Wikipédia est un point de divergence avec l'Encyclopédie de Diderot et d'Alembert
- La "posture de neutralité" consiste à examiner les différents points de vue, sans prendre soi-même position

[a] **Doutreix, M.N. (2010).** *Wikipédia et l'actualité. Qualité de l'information et normes collaboratives d'un média en ligne*. Paris, Presses Sorbonne Nouvelle.

Corpus Élysée

Corpus Élysée: Le corpus des discours présidentiels sous la V° République (1958-2021)

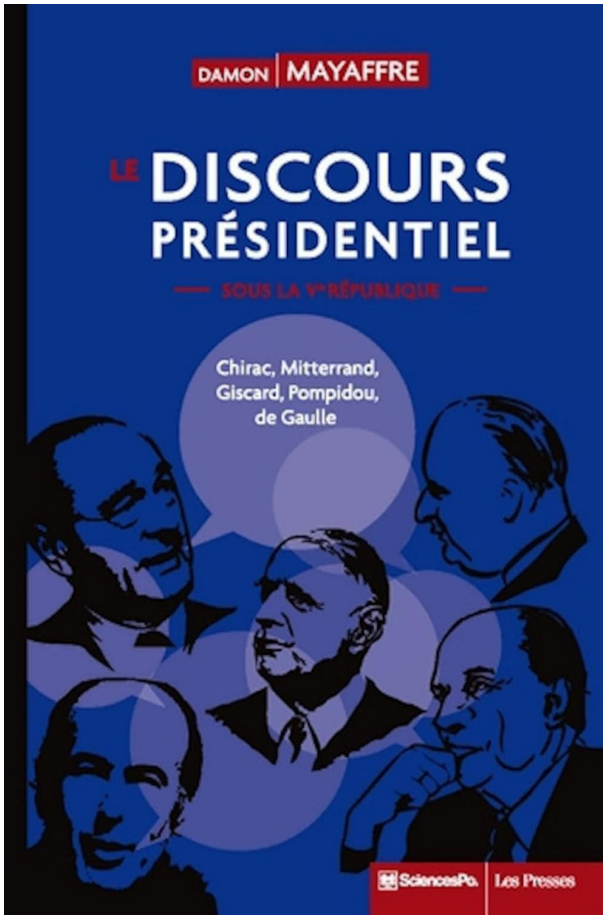


Réalisé par le laboratoire BCL de l'Université Côte d'Azur, ce corpus rassemble les versions écrites de 700 discours des présidents de la Vème république et représente un volume de 2.7 millions de mots.

Variation par président

Président	Années	N discours	N mots
De Gaulle	1958-1968	8	222 876
Pompidou	1968-1974	130	239 355
Giscard	1974-1981	45	369 047
Mitterand	1981-1995	66	697 663
Chirac	1995-2007	116	454 817
Sarkozy	2007-2011	42	266 940
Hollande	2012-2017	48	273 996
Macron	2017-2020	39	318 883
		467	2 843 577

La rhétorique des présidents



Mayaffre (2012)[a]

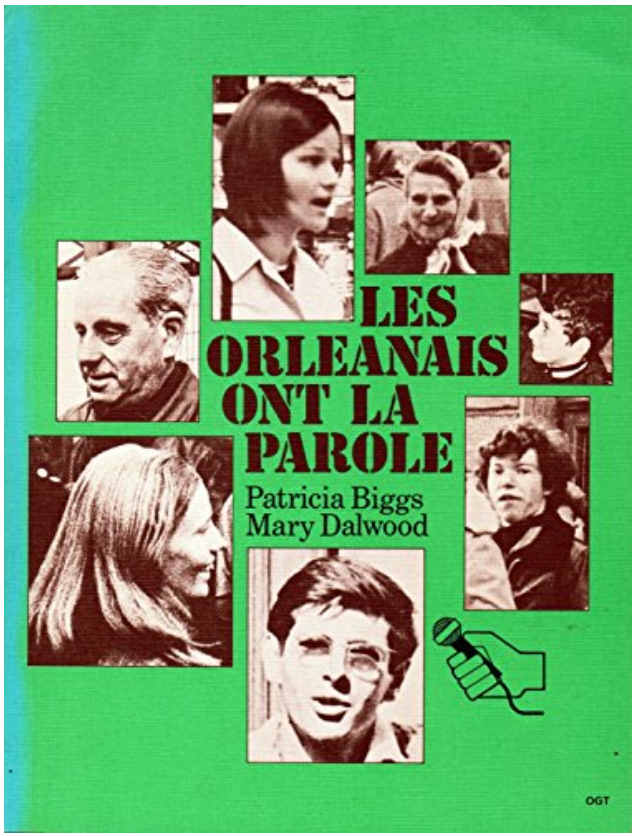
- Transition d'un discours nominal (déterminants, noms et adjectifs) à un discours verbal (pronoms, verbes et adverbes), au tournant des années 1980.
- La posture rhétorique présidentielle de Jacques Chirac procède d'un genre nouveau par rapport à la prose de ses prédécesseurs

---

[a] Mayaffre, D. (2012). *Le discours présidentiel sous la Ve République: Chirac, Mitterand, Giscard, Pompidou, de Gaulle*. Presses de la Fondation de Sciences Politiques.

Corpus ESLO

Enquêtes Sociolinguistiques à Orléans (ESLO)[1]



Très grand corpus de français parlé de 10 millions de mot, les ESLOs sont réalisées par deux enquêtes, l'une initiée en 1966 par des universitaires anglais et l'autre menée dans les années 2010 par les chercheurs du LLL de l'Université d'Orléans, dans la volonté de rendre compte de la diversité des situations de communication

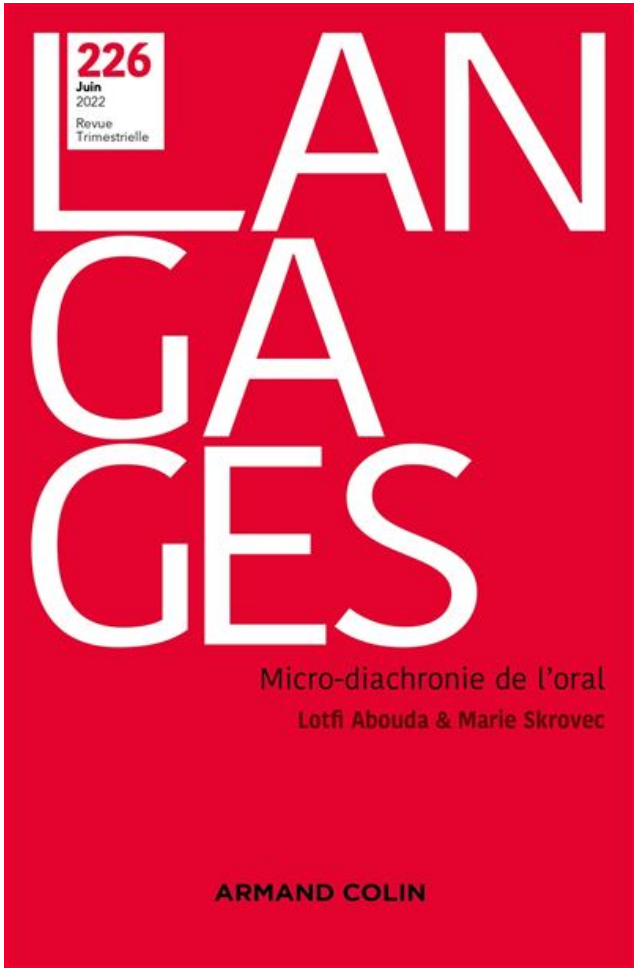
[1] Abouda, Lotfi et Skrovec, Marie. (2018). "Pour une micro-diachronie de l'oral: le corpus ESLO-MD". 6ème Congrès Mondial de Linguistique Française, 46

Variation par situation de communication

ESLO-MD

Situation	Enregistrement	N enregistrements	N mots
Conférences	ESLO1	2	34 638
	ESLO2	5	30 432
Entretiens	ESLO1	30	379 510
	ESLO2	31	452 291
Repas	ESLO1	4	42 373
	ESLO2	8	40 837

D'une génération à l'autre



Abouda et Skrovec (2022)[a]

- L'expression *du coup*, marqueur de conséquence, a explosé en se spécialisant dans l'expression de l'actualisation énonciative
- L'expression *après*, adverbe temporel de postériorité, a développé des emplois non temporels (contrastifs)

[a] Abouda, L. et Skrovec, M. (2022). Michro-diachronie de l'oral. Langages, 2(226)

Variation par situation de communication

ESLO-RAVIOLI

Situation	Enregistrement	N enregistrements	N mots
24H	ESLO2	6	78 568
Ecole	ESLO2	30	143 662
Cinéma	ESLO2	7	42 911
Repas	ESLO2	8	136 350

Corpus BFM

La Base de Français Médiéval[2]

Corpus de français médiéval qui couvre plus de six siècles (du 9ème siècle à la fin du 15ème siècle) de 7 millions de mots, destiné à l'étude des états ancien de la langue française. Développé à l'ENS de Lyon depuis la fin des années 1980 sous l'égide de C. Marchello-Nizia, puis de C. Guillot-Brabance et A. Lavrentiev, pour le versant TXM.

[2] Alexei Lavrentiev and Céline Guillot-Barbance (2024) "La BFM 2022 : un corpus pour les recherches diachroniques en français médiéval et au-delà", *Corpus*, 25

Variation par période chronologique

Période	Siècle	N textes	N mots
Très ancien français	9-11ème	7	12 009
Ancien français	12ème	63	1 595 623
	13ème	60	2 069 099
Moyen français	14ème	32	1 734 284
	15ème	57	1 917 820

L'histoire d'un objet culturel : la phrase

Siouffi (2020)[a]

- Les premiers textes de la langue française sont chantés, psalmodiés, lus à voix haute devant un public
- L'écrit argumentatif s'accélère au 14ème siècle pour répondre aux besoins de la communication politique mais aussi à la diffusion du savoir.

---

[a] Siouffi, G. (2020) (dir.), *Une histoire de la phrase française des Serments de Strasbourg aux écritures numériques*, Arles, Actes Sud.