Université Grenoble Alpes

Laboratoire d'Informatique de Grenoble

Equipe Systèmes d'Information - Ingénierie et Modélisation Adaptables (SIGMA)

Bibliothèque du Français Moderne (BEFM)

Description du corpus

« Le vocabulaire de Victor Hugo »

Janvier 2021

Le « corpus Hugo » en ligne sur la BEFM, est présenté en détail en annexe2. Il se compose de 8 recueils de poésie – dont les trois *Légende des siècles* --, 5 romans – dont *Notre-Dame de Paris* et les *Misérables*-, 8 pièces de théâtre et 22 ans de correspondance (dont la période de son exil qui est la plus fournie en lettres), soit au total plus de 2 millions de mots répartis en 156 fichiers.

Origine des textes électroniques :

- Association des Bibliophiles Universels (ABU) : les *Contemplations*, les *Misérables* (tomes 1 et 2), *Notre-Dame de Paris*,

- Dominique Labbé à partir de l'édition originale en ligne sur Gallica : *Marion Delorme*, *Marie Tudor*.

- Wikisource : 1793, *l'Art d'être grand-père*, les *Châtiments*, *Correspondance*, *Cromwell*, *Hernani*, *l'Homme qui rit*, *la Légende des siècles*, *Lucrèce Borgia*, *les Misérables* (tomes3 et suivants), *Odes et ballades*, *Ruy Blas*, les *Travailleurs de la mer*

Dominique Labbé a assuré la correction et la standardisation orthographiques, le balisage et l'étiquetage de ces textes (de 1999 à 2020).

Pour tout détail, écrire à Dominique Labbé (dominique.labbe@umrpacte.fr)

Annexe 1

Etudes statistiques sur l'œuvre de Hugo

Trois études sur un corpus assez proche mais sur les formes graphiques :

Brunet Etienne (1988a). *Le vocabulaire de Victor Hugo*. Paris : Champion-Slatkine, 1988.

Brunet Etienne (1988b). Hugocentric Tendencies or Can One Approach Hugo Counting Words. *Literary and Linguistic Computing*. 1988, 2, p. 79-9.

Brunet Etienne (1988c). La structure lexicale dans l'oeuvre de Hugo. In Labbé Dominique, Philippe Thoiron et Serant Daniel (dir.). *Etudes sur la richesse et la structure lexicales*. Paris : Champion-Slatkine, p. 24-42.

Trois études sur les textes étiquetés :

Labbé Cyril, Labbé, Dominique & Urien Frédéric (2020). La bibliothèque en ligne du français moderne. Le vocabulaire de V. Hugo. *Semaine Data-SHS. "Traiter et analyser des données en sciences humaines et sociales"*. Plateforme Universitaire de Données (TIR-PROGEDO). Université de Grenoble-Alpes. 7-11 décembre 2020. https://www.researchgate.net/publication/346898988

Labbé Dominique (2014). Identification de l'auteur d'un texte (Hugo, Lamartine, Musset et Vigny). Conférence invitée au séminaire *L'œuvre et son auteur : problèmes d'attribution*. Lille : Université de Lille-Nord de la France, 21 mai 2014. https://www.researchgate.net/publication/278778502

Labbé Cyril, Labbé Dominique (2013). Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant. In Banks David. *Le texte épistolaire du XVIIe siècle à nos jours*. Paris : L'Harmattan, 2013, p. 53-85. https://www.researchgate.net/publication/39063948

Annexe 2

Le corpus Hugo (janvier 2021)

	Dates	Mots	Vocables
Correspondance	1849-1870	292 153	9 925
Poésie			
Odes et ballades	1828	57 957	4 557
Contemplations	1830-1852	91 887	5 893
Châtiments (les)	1853	55 452	5 851
Légende des siècles	1859-1883	215 759	10 278
L'Art d'être grand père	1877	31 708	3 870
Total Poésie		**452 763**	**13 710**
Romans			
Notre Dame de Paris	1831	185 469	10 682
Misérables (les)	1862	564 234	17 146
Travailleurs de la mer (les)	1866	141 133	9 738
Homme qui rit (l')	1869	199 971	12 326
1793	1874	123 712	8 382
Total Romans		**1 214 519**	**26 478**
Théâtre			
Cromwell	1827	81 871	6 858
Hernani	1830	19 577	2 304
Marion Delorme	1831	19 641	2 295
Le Roi s’amuse	1832	15 986	1 948
Marie Tudor	1 833	22 129	1 881
Lucrèce Borgia	1 833	19 778	2 186
Ruy Blas	1 838	24 190	2 984
Torquemada	1869	20 036	2 725
Total Théâtre		**223 208**	**9 622**
Total général		**2 182 643**	**32 484**