La base Frantext

Plan du cours

1 Présentation

À partir des années soixante l'apparition d'ordinateurs de plus en plus puissants commence à faciliter l'analyse linguistique des textes. L'année 1964 marque le véritable coup d'envoi des travaux sur corpus, avec la constitution du premier corpus informatisé par une équipe de chercheurs de l'université Brown aux États-Unis. En France, une dizaine d'années plus tard, on commence à rassembler la base textuelle Frantext.

En septembre 2019, la base de données Frantext comportait 5418 références, soit 254 millions de mots. Sa particularité est de coupler un corpus échantillonné du IXe au XXIe siècle et un outil de recherche performant. Frantext contient entre autres une importante proportion de textes modernes et contemporains. Elle contient 10% de textes dits « scientifiques » et techniques et 90% de textes considérés comme « littéraires » regroupant tous les genres : romans, mémoires, autobiographie, journaux personnels, théâtre, poésie, essais. Ce corpus comprend des œuvres françaises, mais aussi francophones.

La base Frantext est avant tout un outil d'analyse textuelle et ne permet pas d'afficher une œuvre dans son intégralité. Elle permet de faire des recherches à différents niveaux : retrouver une citation exacte et son auteur, rechercher les occurrences d'un terme ou d'une expression dans un corpus d'œuvres sélectionnées ou sur tous les textes de la base, calculer des fréquences d'usage, rechercher des listes de mots par exemple.

2 Fonctionnement

Notez d'abord que, pour accéder à Frantext intégral, vous devez passer par le portail du Service commun de documentation de l'université (La doc en ligne › Langues et linguistique).

Le principe de fonctionnement de Frantext est simple [1] : l'utilisateur peut consulter cette base grâce à un certain nombre de paramètres plus ou moins avancés et peut ainsi croiser des critères tels que le genre, l'œuvre, la date (plus

1. Pour une première prise en main, je vous renvoie au lien « Frantext pour les nuls » que vous trouverez dans ce cours.

ou moins précise), l'édition ou bien encore la graphie d'un lemme. Frantext élabore également des statistiques de fréquence (par décennie, siècle, etc.).

Le corpus de travail ainsi établi (une œuvre, un genre, un siècle, ou l'ensemble du corpus soit plus de 3 600 textes), l'utilisateur peut alors interroger la base de données et rechercher les occurrences d'un mot, d'une séquence, les co-occurrences de mots ou de séquences de mots mais il peut aussi indiquer les flexions verbales, adjectivales, substantivales, et ce grâce aux grammaires paramétrables. Il est aussi possible d'étudier le voisinage d'un mot recherché. En somme, le degré de complexité de la demande dépend uniquement de l'interrogation formulée initialement.

3 L'intérêt de Frantext

Linguistes, historiens, philosophes peuvent utiliser cette base afin de mieux définir le sens des mots et concepts utilisés par un courant, une école, un auteur ou bien simplement effectuer une recherche transversale.

En consultant la base, l'étudiant ou le chercheur peut notamment suivre l'évolution sémantique d'un mot ou d'un groupe de mots depuis les premiers temps de la langue française jusqu'en français moderne et ainsi circonscrire avec précision les néologismes sémantiques au fil des siècles, appréhender les différentes acceptions d'un mot en contexte au travers de milliers de textes (littérature, politique, etc.) qui constituent la base.

4 Exemples d'application

4.1 *Sido* de Colette

L'analyse qui suit est tirée de l'article de Danielle Bouverot, « La base Frantext au service de la stylistique » [2]. Dans ce recueil de souvenirs d'enfance écrit par Colette, le premier mot lexical qui apparaît, dès le 33e rang, alors que les premiers vocables sont toujours des mots grammaticaux, est mère,

2. In : *L'Information Grammaticale*, N. 70, 1996. p. 38-42.

de fréquence 94, suivi de peu, au 49e rang, par père 59, puis par enfant 42, aîné 26, frère 24, fille 20, cadet 11 , mari 9, sœur 8, si nous rassemblons le singulier et le pluriel, pour les 500 premiers mots de la liste, dont la fréquence est au moins 4 : il s'agit tout naturellement du vocabulaire élémentaire de la famille, ce qui n'étonne pas le lecteur de ce genre de texte. Presque à égalité survient un vocabulaire des parties du corps, attendu pour une description des personnages [3]. Mais d'autres thèmes s'entrecroisent, celui des couleurs, celui de la végétation, lié au jardin de Sido. Ces quatre champs lexicaux ont en commun d'être formés de mots très courants, qui passent inaperçus dans leur banalité.

Par ailleurs, en examinant le contexte où sont employés ces termes, on note l'imbrication de certains champs lexicaux. Le vocabulaire des couleurs est certes justifié par la description et le portrait, mais il révèle aussi la psychologie d'un personnage : « Car <Sido> aimait au jardin le rouge, le rose, les sanguines filles du rosier, de la croix- de-Malte, des hortensias et des bâtons de Saint Jacques, et même le coqueret-alkékenge, encore quelle accusât sa fleur, veinée de rouge sur pulpe rose, de lui rappeler un mou de veau frais » (p. 26). D'autre part, la polysémie, attestée dans la langue, associe une fleur et une couleur, par exemple, avec rose et lilas ; ainsi « un corsage lilas, un buisson de cheveux crêpelés, d'un rose de cuivre, éclairèrent le haut de la rue » ; Colette la complète dans la même page par une métonymie rare, où un personnage féminin, au prénom d'ailleurs évocateur, se confond avec un végétal : « C'est Flore Chebrier qui rejoint son père, dit mon frère aîné quand l'or et le lilas s'éteignirent au bas de la rue » (p. 158). Les figures d'analogie contribuent au tissage des thèmes : l'être humain et la végétation s'associent ; les « cheveux crêpelés » de la même Flore deviennent peu après un « bouquet de cheveux roux » (p. 160). Par une sorte de contamination, le thème de la famille gagne le règne végétal, avec « les sanguines filles du rosier » (p. 26) ou une « famille de bulbes » (p. 41).

3. Yeux 38 (à rapprocher de regard 15, vois 9, vue 6), puis main 21, tête 19, pied 13, bras 8, bouche 7, lèvres, menton, nez, sourcils 6, barbe 5, front, oreilles 4.

Conclusion La consultation du contexte grâce à Frantext permet de proposer une explication stylistique fine qui rend plus perceptible le style d'un écrivain.

4.2 L'étude de l'adjectif « deshonneste »

En français, l'adjectif deshonneste est attesté depuis le XIIIe siècle avec le sens de « affreux ». Denis Foulechat l'emploie dans sa traduction française du *Polycratique* à la fin du XIVe siècle. Quelle est l'histoire de ce terme [4] ? C'est ici que la base de données Frantext joue un rôle essentiel en termes de prospection quantitative. Le nombre d'occurrences répertoriées parle assez nettement : 186 attestations de l'adjectif couvrant tous les siècles prouvent son ancrage sémantique. Voici leur répartition par siècle et par graphie :

Tableau 1. ***Répartition des occurrences de* deshonneste *par siècle et par graphie dans FRANTEXT***

	XVIe siècle	*XVIIe siècle*	*XVIIIe siècle*	*XIXe siècle*	*XXe siècle*
Deshonneste	34	65	2	0	0
Deshonnête	1	2	8	2	0
Déshonnête	0	4	30	24	14

Les attestations des XIXe et XXe siècles sont suffisamment nombreuses pour affirmer que le terme est vivant dans la langue. Mais une étude approfondie des occurrences permet aussi de le circonscrire presque exclusivement dans le champ littéraire. Ainsi le trouve-t-on chez Zola, Verlaine, les frères Goncourt, Huysmans, Alain, Camus, Gide, Duhamel, Genevoix etc. Le dictionnaire général Larousse du XXe siècle donne une définition du terme qui valide son ancrage lexical restreint : « Déshonnête, adj. LITT. Contraire à la morale, à la pudeur. Geste déshonnête (synonyme : inconvenant, indécent) ». La définition conserve le sème de la moralité mais le restreint à un emploi littéraire dont la langue commune est écartée. Une telle désaffection n'était pas notée chez les lexicographes antérieurs même si les définitions qu'ils donnaient du mot

4. Pour répondre à cette question, nous nous appuyons sur l'analyse d'Olivier Bertrand, « De l'usage de la base de données Frantext en sciences humaines et sociales », *Histoire & mesure*, XVIII - 3/4 | 2003, 375-387.

restaient fortement dans le domaine moral puisqu'on lit dans le dictionnaire Littré : « Deshonnête, adj. Qui est contre l'honnêteté ou la pudeur. Pensées, paroles, actions, manières déshonnêtes. SYN. MALHONNÈTE. Déshonnête est contre la pureté, la pudeur ; malhonnête est contre la civilité, et, quelques fois, contre la bonne foi. »

Dans cette définition, on ne trouve aucune mention d'un quelconque archaïsme dans l'emploi de cet adjectif. En revanche, il a bien un sème moral, ce que traduit l'opposition sémantique relevée par Littré entre déshonnête et malhonnête. Si l'on compare leur nombre d'occurrences dans la base de données Frantext, il semble que les courbes quantitatives, en termes de fréquence absolue, se présentent de manière inverse :

Tableau 2. Évolution de déshonnête/malhonnête en nombre d'occurrences dans FRANTEXT

	XVI^e siècle	*XVII^e siècle*	*XVIII^e siècle*	*XIX^e siècle*	*XX^e siècle*
Déshonnête	35	71	40	26	14
Malhonnête	0	25[4]	174	104	159

Conclusion Alors que « malhonnête » est d'emploi assez peu courant avant le XVIII^e s., l'adjectif multiplie ses attestations au cours des XIX^e et XX^e s. En revanche, « déshonnête » connaît un déclin certain à partir du XVIII^e s. Il semblerait, à la consultation des dictionnaires postérieurs au XIX^e s., que l'adjectif « malhonnête » a repris le sème moral contenu jusqu'alors dans « déshonnête » devenu moribond sauf dans le monde littéraire, plus conservateur, qui entend encore marquer la distinction sémantique entre les deux adjectifs.

Vous savez maintenant comment utiliser la plupart des fonctionnalités de la base Frantext. il reste à étudier la fonction de « recherche avancée » qui permet d'utiliser des expressions régulières et des requêtes CQL.

5 Les expressions régulières

Le moteur de recherche de Frantext permet l'utilisation d'expressions régulières – appelées également expressions rationnelles – pour rechercher des

suites de caractères selon des motifs et des formules logiques. Les expressions régulières peuvent être utilisées dans les recherches avancées, de fréquences, de co-occurrences et de voisinages, ainsi que dans les listes de mots.

5.1 Opérateurs

Les opérateurs permettent de spécifier le type de caractères à rechercher [5].

Opérateur	Description	Exemple de recherche avancée	Exemple de résultats séparés par des virgules
()	Groupe de caractères d'une expression	"(nuit)"	nuit
|	Choix entre plusieurs alternatives	"jour|nuit"	jour, nuit
.	N'importe quel caractère	"n.it"	nuit, nait...
\	Interpréter littéralement un opérateur	"\."	.
[]	Un des caractères entre crochets	"[bp]eau"	beau, peau
[^]	Tout caractère hormis ceux entre crochets	"n[^u]is"	nais, nois
[a-z]	Un intervalle composé de caractères alphabétiques de a à z, en minuscules et sans diacritiques	"[a-z]ait"	fait, sait, lait...
[0-9]	Un intervalle composé de chiffres de 0 à 9	"[0-9]"	0, 1, 2, 3, 4, 5, 6, 7, 8, 9

5.2 Quantificateurs

Les quantificateurs permettent de spécifier le nombre de caractères à rechercher.

6 Les expressions CQL

CQL est l'acronyme de *Corpus Query Language* : il s'agit d'un langage d'expression de requêtes. Une expression CQL est une chaîne de caractères exprimant un motif linguistique – un mot, ou une suite de mots – défini en fonction de formes graphiques, de formes lemmatisées ou de catégories

5. Seuls les éléments surlignés sont à connaître.

Quantificateur	Description	Exemple de recherche avancée	Exemple de résultats séparés par des virgules
?	**Zéro ou une fois** le caractère ou groupe qui précède	"nu(it)?"	nuit, nu
*	**Zéro ou plusieurs fois** le caractère ou groupe qui précède	"cré*e"	cre, crée, créée
+	**Une ou plusieurs fois** le caractère ou groupe qui précède	"cré+e"	crée, créée
{n}	**Exactement n occurrences** de l'expression précédant les accolades. La valeur de **n** est limitée à 32.	"(ha){2}"	haha
{n,n}	**Exactement n occurrences** de l'expression précédant les accolades. La valeur de **n** est limitée à 32.	"(ha){2,2}"	haha
{n,m}	**Entre n et m occurrences** de l'expression précédant les accolades. Les valeurs de **n** et **m** sont limitées à 32.	"(ha){2,3}"	haha, hahaha
{n,}	**Au moins n occurrences** de l'expression précédant les accolades. La valeur de **n** est limitée à 32.	"(ha){2,}"	haha, hahaha, hahahaha...

grammaticales. Les expressions CQL peuvent être combinées, utiliser des expressions régulières, et tenir compte de variante d'écritures.

6.1 Les mots réservés

Utilisez les crochets et les mots réservés word, lemma et pos [6] pour effectuer des recherches spécifiques.

Expression	Description	Exemple de résultats
[word="bonheur"]	La forme graphique	bonheur
[lemma="aimer"]	Toutes les formes (conjuguées ou non) du verbe	aime, aimer, aimait, etc.
[pos="VINF"]	Tous les verbes à l'infinitif.	être, faire, avoir, etc.

6. Pour la syntaxe de pos, on notera : NC = nom commun, ADJ, ADV, DET = déterminant, PRE = préposition et VINF = verbe à l'infinitif.

La recherche d'une forme graphique étant la recherche la plus fréquente, une forme simplifiée est disponible. Il suffit de saisir la forme graphique entre guillemets (veillez à bien utiliser les guillemets doubles droits).

6.2 Les recherches combinées

Il est possible de combiner des requêtes à l'aide des opérateurs booléens logiques ET &, OU | et NON !.

Expression	Description
[word="grand" & pos="NC"]	Toutes les occurrences de la forme graphique grand utilisée comme nom commun
[lemma="grand" & pos="ADJ"]	Toutes les occurrences du lemme grand utilisé comme adjectif
[word="grand" | word="petit"]	Toutes les formes graphiques correspondant à grand ou petit
[lemma="grand|petit"]	Variante d'écriture utilisant les expressions régulières pour trouver tous les lemmes correspondant à grand ou petit
[word="grand" | lemma="petit"]	Toutes les occurrences de la forme graphique grand ou du lemme petit
[lemma="grand" & !(pos="NC")]	Toutes les occurrences du lemme grand lorsqu'il n'est pas utilisé comme nom commun
[lemma="grand" & pos!="NC"]	Toutes les occurrences du lemme grand lorsqu'il n'est pas utilisé comme nom commun (variante d'écriture. Le point d'exclamation doit toujours être collé au signe égal)