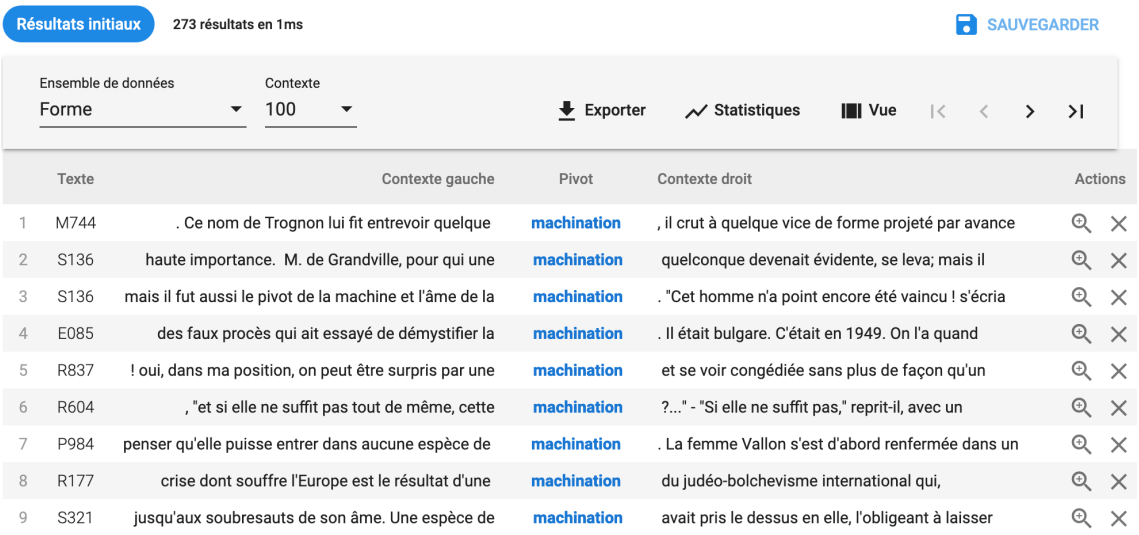
Manipulations avec « Frantext »

Frantext est une base de donnée textuelle, qui regroupe un important corpus de textes français, du IX^e^ au XXI^e^ siècle, saisis sur support informatique. Le corpus est constitué d'environ 5500 œuvres, dont environ 90 % d'œuvres littéraires et 10 % d'ouvrages techniques.

Une interface de consultation permet de rechercher, dans le corpus ou dans une partie, diverses séquences et leur contexte. Par exemple, si on recherche toutes les occurrences du mot *machination*, on obtiendra une suite d'extraits comme le suivant :



	Texte	Contexte gauche	Pivot	Contexte droit	Actions
1	M744	. Ce nom de Trognon lui fit entrevoir quelque	machination	, il crut à quelque vice de forme projeté par avance	
2	S136	haute importance. M. de Grandville, pour qui une	machination	quelconque devenait évidente, se leva; mais il	
3	S136	mais il fut aussi le pivot de la machine et l'âme de la	machination	. "Cet homme n'a point encore été vaincu ! s'écria	
4	E085	des faux procès qui ait essayé de démystifier la	machination	. Il était bulgare. C'était en 1949. On l'a quand	
5	R837	! oui, dans ma position, on peut être surpris par une	machination	et se voir congédiée sans plus de façon qu'un	
6	R604	, "et si elle ne suffit pas tout de même, cette	machination	?..." - "Si elle ne suffit pas," reprit-il, avec un	
7	P984	penser qu'elle puisse entrer dans aucune espèce de	machination	. La femme Vallon s'est d'abord renfermée dans un	
8	R177	crise dont souffre l'Europe est le résultat d'une	machination	du judéo-bolchevisme international qui,	
9	S321	jusqu'aux soubresauts de son âme. Une espèce de	machination	avait pris le dessus en elle, l'obligeant à laisser	

Cette interface est accessible via Internet (aux organismes abonnés), avec un navigateur, et c'est cette interface qui fait l'objet des manipulations d'aujourd'hui.

Pour accéder à Frantext, avec le navigateur de votre choix (`firefox` par exemple), on passe par `Virtuose+`. Une fois connectés, allez dans `Mon compte` puis `Mes bases`, recherchez `Frantext`, puis dans la nouvelle fenêtre, choisissez « Frantext intégral ».

Manipulations

1. *Sélection du corpus.* Par défaut, on travaille avec l'intégralité des textes, dont les plus anciens sont dans une langue assez éloignée de la nôtre (on y trouve les deux vers « *Et ma despoille me gardés, Jo vos lais ma vie et ma mort.* »). Il est donc préférable de choisir un corpus plus homogène, par exemple le corpus prédéfini Classique qui regroupe 1103 œuvres (on verra plus loin comment définir un corpus spécifique). Menu Corpus.

2. *Recherche.* Menu « Recherche ». Faites une recherche simple pour commencer, en choisissant un mot pas trop fréquent, par exemple le mot *franchement*.
Sur la page de résultats, explorer les différentes options offertes : Statistiques, Vue, Ensemble de données.

3. Toujours dans le même corpus, rechercher les occurrences de *ne plus*. Est-ce que ces

expressions sont correctement catégorisées par Frantext ? Quelles sont les catégories syntaxiques qui peuvent apparaître à la suite de *ne plus* ?

4. *Recherche de co-occurrences.* Menu « Recherche/Co-occurrences ». Essayons maintenant de rechercher la séquence *ne... plus* où les deux peuvent être séparés. En examinant dans chaque cas les 25 premiers résultats, déterminez quelle est la meilleure façon de spécifier la recherche de co-occurrences.

5. *Variantes flexionnelles.* Menu « Recherche/Avancée ». Pour repérer toutes les variantes flexionnelles d'un mot (c'est-à-dire pour un verbes ses différentes formes conjuguées, pour un nom/adjectif ses différentes formes flêchies), on peut utiliser le langage CQL qui permet de rechercher par exemple toutes les formes du verbe *refuser* au moyen de la requête `[lemma = "refuser"]`.

6. *Utilisation des catégories.* Menu « Recherche/Avancée ». Il arrive que le même lemme corresponde à deux catégories grammaticales (parties du discours, engl. *POS-tag*). Par exemple le mot *aveugle* peut être aussi bien un nom qu'un adjectif. Une requête utilisant le code `pos` permet de choisir la catégorie que l'on cherche : par exemple `[lemma = "aveugle" & pos = "ADJ"]`. Quel est l'emploi majoritaire de *aveugle* dans ce corpus ? Nom, Adjectif ?

7. *Complémentaire.* Le langage de requête permet aussi d'exprimer une négation. Avec la notation `[pos != "ADV"]` on peut trouver les occurrences qui ne sont pas de la catégorie « adverbe ». Quelles sont les autres catégories qui sont associées au mot *plus* en plus de la catégorie "Adverbe" ? Est-ce que vous pensez que ces occurrences sont correctement catégorisées ?

8. *Listes de mots.* Menu « Liste de mots ». Pour faciliter les recherches, il est possible de créer et sauvegarder des listes de mots. Par exemple, en français, les linguistes établissent la liste des semi-négations (les mots qui se combinent avec *ne* pour former des phrases négatives) comme comprenant (au moins) : *personne*, *rien*, *nul*, *aucun*, *jamais*, *plus*, *nulle part*. Créer une liste des semi-négations du français et chercher les phrases qui comportent à la fois le mot *ne* et une semi-négation.

9. *Récupération des résultats* Menu « Exporter » dans la fenêtre des résultats. Il est possible de récupérer les extraits trouvés en les exportant au format csv, txt ou xml, et en incluant les méta-données pertinentes. Vérifez que vous pouvez exporter les résultats dans un format lisible.

10. *Annotation.*
On va tenter de répondre à la question suivante : lorsque 2 semi-négations apparaissent dans une même phrase (comme dans *Amine n'a rien dit à personne*), quelle est la proportion des cas où l'interprétation est « mono-négative » vis-à-vis des cas où elle est « bi-négative » ? Pour répondre à cette question, on va travailler sur le corpus du 20e siècle, et extraire 25 occurrences de phrases comportant deux semi-négations. Ces extraits vont être sauvegardés dans un fichier, et pour chaque exemple, on va l'annoter selon qu'il est mono-négatif, bi-négatif, ou non-annotable.