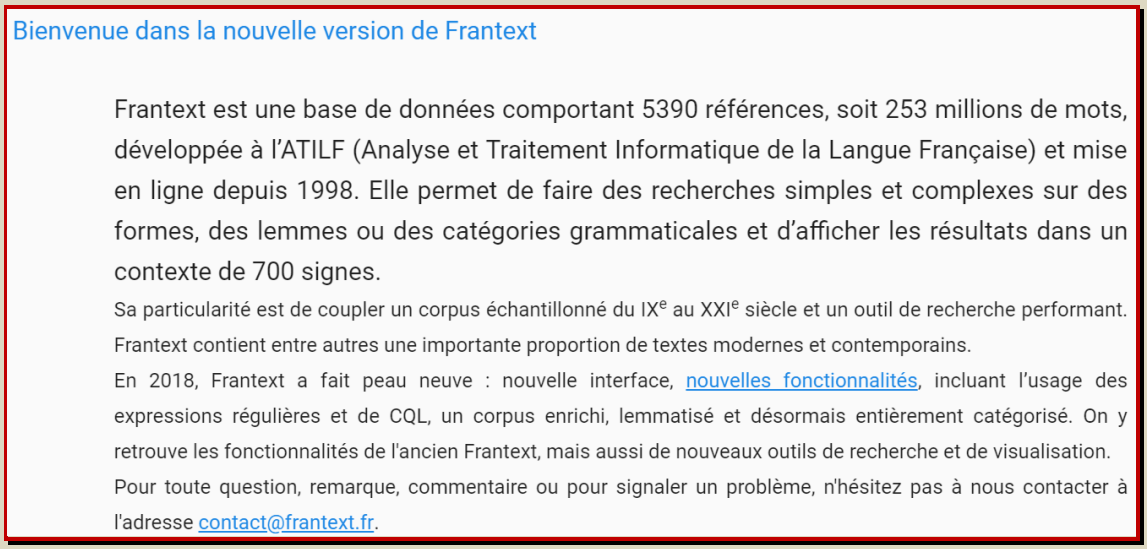
Jacques FRANÇOIS
Université de Caen & CRISCO, EA 4255

La base de données textuelles FRANTEXT-2 : Comment en tirer parti dans les études linguistiques et littéraires



Bienvenue dans la nouvelle version de Frantext

Frantext est une base de données comportant 5390 références, soit 253 millions de mots, développée à l'ATILF (Analyse et Traitement Informatique de la Langue Française) et mise en ligne depuis 1998. Elle permet de faire des recherches simples et complexes sur des formes, des lemmes ou des catégories grammaticales et d'afficher les résultats dans un contexte de 700 signes.

Sa particularité est de coupler un corpus échantillonné du IX^{e} au XXI^{e} siècle et un outil de recherche performant. Frantext contient entre autres une importante proportion de textes modernes et contemporains.

En 2018, Frantext a fait peau neuve : nouvelle interface, nouvelles fonctionnalités, incluant l'usage des expressions régulières et de CQL, un corpus enrichi, lemmatisé et désormais entièrement catégorisé. On y retrouve les fonctionnalités de l'ancien Frantext, mais aussi de nouveaux outils de recherche et de visualisation.

Pour toute question, remarque, commentaire ou pour signaler un problème, n'hésitez pas à nous contacter à l'adresse contact@frantext.fr.

La base de données textuelles FRANTEXT a été créée initialement comme un outil interne à l'*Institut National de la Langue Française* (CNRS Nancy, INALF), en vue de fournir aux rédacteurs du *Trésor de la Langue Française* (en 16 volumes et en ligne sur le site du CNRTL) une grande quantité d'occurrences de tous les mots destinés à figurer dans ce dictionnaire. Après la publication de la version informatisée du dictionnaire, FRANTEXT a été ouvert au public, toujours sur le site du CNRTL, malheureusement sur abonnement, contrairement au *British National Corpus* en Grande-Bretagne ou au Corpus de référence COSMAS II de l'*Institut für Deutsche Sprache* à Mannheim, en Allemagne. La plupart des universités européennes ont un abonnement institutionnel à la base de données. L'abonnement individuel annuel se monte à 42€ TTC.

Public concerné : Chercheurs, enseignants-chercheurs, doctorants.

Durée de l'abonnement : 1 an. L'abonnement doit être renouvelé chaque année. L'abonnement est ouvert soit pour l'**année civile**, du 1er janvier au 31 décembre, soit pour l'**année universitaire**, du 1er octobre au 30 septembre. Précisez la période souhaitée lors de votre commande.

Tarif annuel : 35 € HT (42 € TTC). C'est à l'abonné et non à l'institution de rattachement (université, ...) d'en assurer le règlement. Après réception de la commande ou du paiement, les informations de connexion vous seront délivrées par courrier électronique. Le coût est forfaitaire.

Durant l'été 12018, le mode d'emploi de la base, appelée désormais FRANTEXT-2, a été profondément révisé. L'objectif de ce didacticiel est d'initier les internautes désireux d'explorer la langue française – essentiellement littéraire – d'aujourd'hui ou du passé (jusqu'au XII^{e} siècle) à quelques aspects majeurs de son exploitation et du transfert et traitement du résultat des recherches dans un tableur. Il se

composera de quatre chapitres délivrés successivement en mai et juin 2019 sur le site www.interlingua.fr ⇨

Chap.1 : Sélection un corpus et recherche "simple" ;
Chap.2 : Recherche "assistée" et recherche "avancée";
Chap.3 : Transfert des résultats dans un tableur et leur exploitation;
Chap.4 : Trois études de cas.

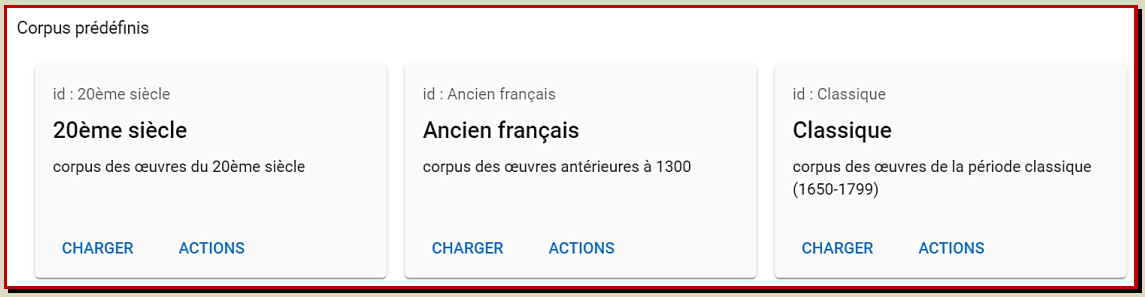
Chap.1 : Sélection d'un corpus et recherche "simple"

1.1. *Les corpus prédéfins et les corpus personnels*

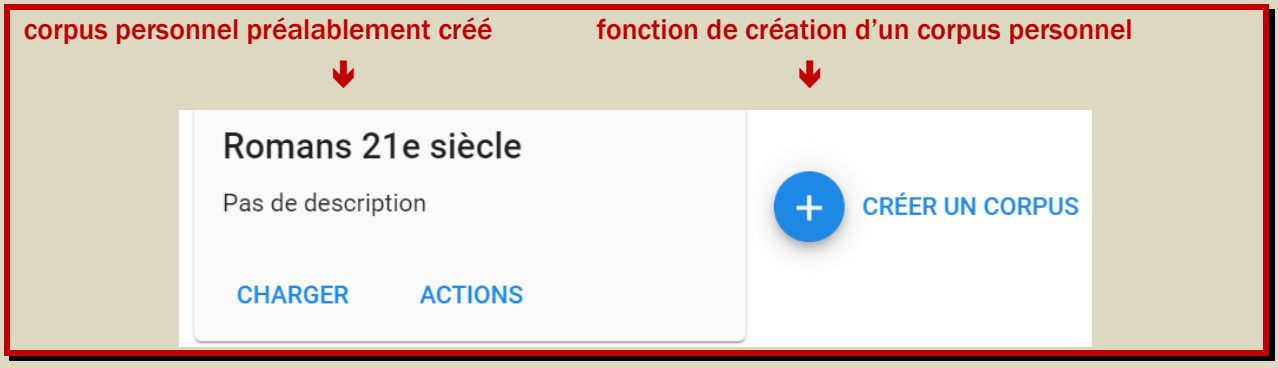
Une fois que vous êtes connecté, FRANTEXT-2 vous propose trois variantes de la base :

La base de démonstration permet de s'initier gratuitement aux modalités des recherches, mais son corpus étant limité à 40 textes, nous passerons directement à la version complète, FRANTEXT intégral, dans laquelle vous devrez sélectionner un corpus :

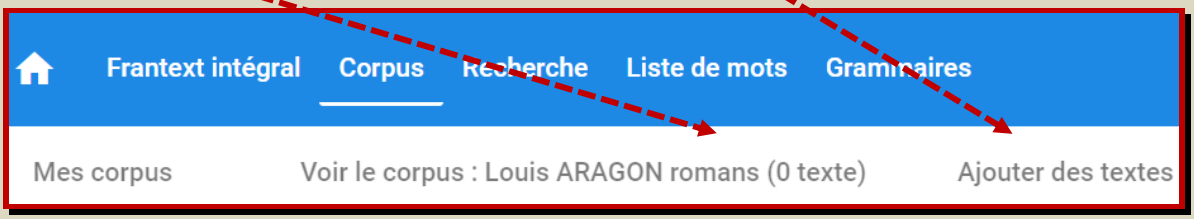
Dix corpus prédéfinis sont proposés, dont voici les trois premiers :



Si aucun de ces dix corpus ne correspond à vos besoins, vous pouvez créer vous-même un corpus :



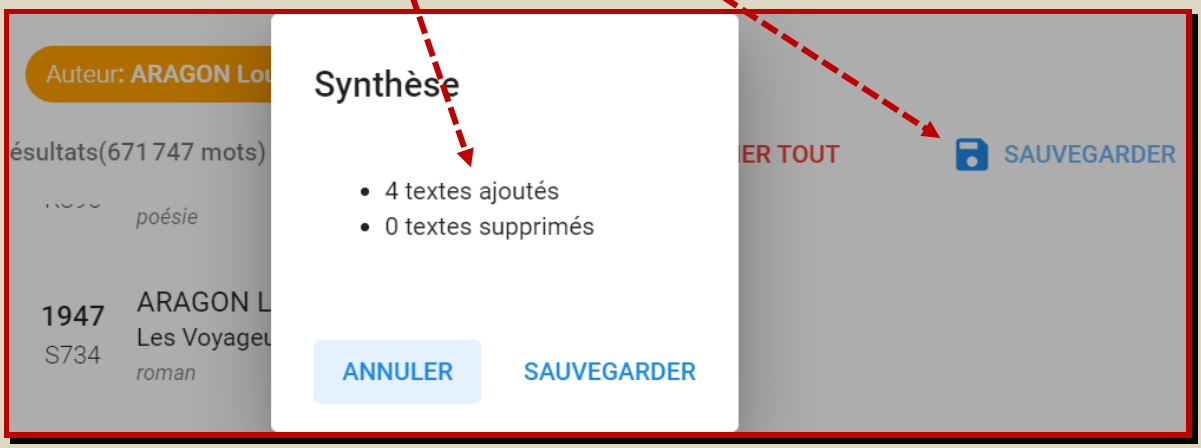
Si vous saisissez comme titre de votre corpus « Louis ARAGON romans », vous créez un corpus initialement vide : il vous appartient donc d' ajouter des textes



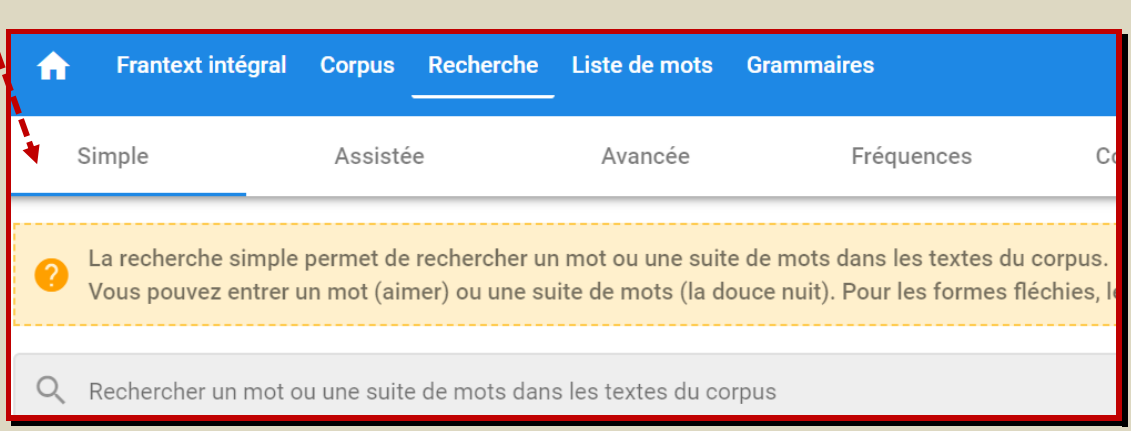
14 textes sont enregistrés sous le nom de Louis ARAGON.

Le genre littéraire de chaque texte est indiqué, ainsi que sa date de publication et son volume en nombre de mots :

Vous pouvez sélectionner les 14 textes, puis décocher les poésies et les essais, ou procéder en sens inverse. Vous enregistrez 4 romans qu'il faut sauvegarder :



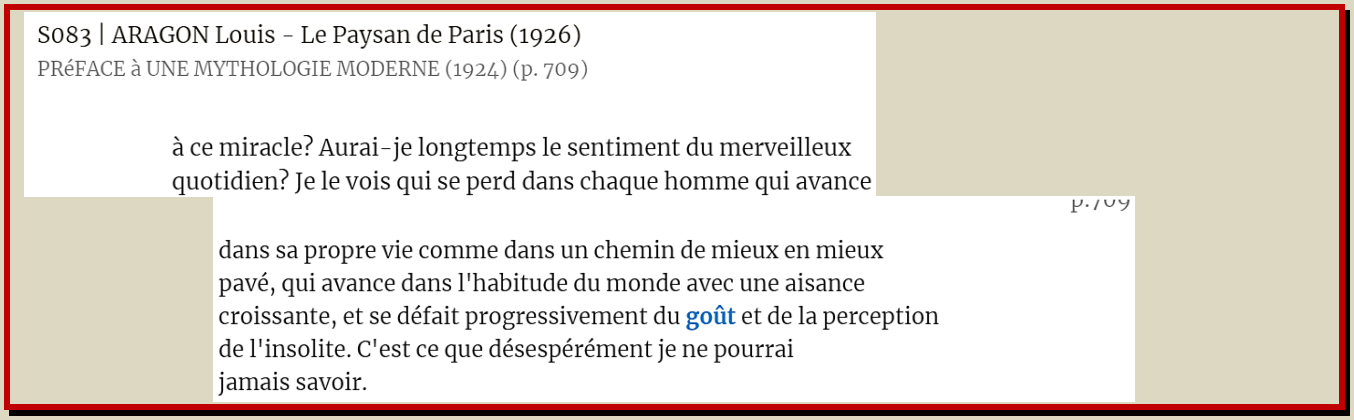
Désormais, vous pouvez effectuer une recherche. Le mode élémentaire est celui de la recherche simple :



Le recherche des occurrences de « goût » dans ces 4 romans délivre 121 occurrences, celle du pluriel « goûts » seulement 9. Avec un rapport de 1 à 13 entre le pluriel et le singulier, on peut déjà en déduire qu'Aragon s'intéresse moins à la variété des goûts qu'au goût des choses, des sensations, etc.

Les 121 résultats pour le singulier « goût » se présentent en priorité sous la forme d'un concordancier en trois champs : contexte gauche (200 caractères au maximum) / pivot (= chaîne de caractères recherchée) / contexte droit (même nombre de caractères). Il est possible de changer de vue au profit du contexte afin de savoir de quelle œuvre et à quelle page provient chaque occurrence :

Vue du contexte de la première occurrence délivrée :



À partir de la vue sur le concordancier on peut également sélectionner une « vue compacte » moins détaillée (sans pagination) mais contenant 500 caractères :

Le résultat de la recherche doit être soit sauvegardé dans FRANTEXT 2, soit plutôt exporté. Plutôt que le format « txt », il vaut mieux sélectionner le format « csv » qui prépare la répartition des trois champs dans trois colonnes d'un tableur. L'export des résultats est modulable. Dans l'export spécifié ci-après, Le nombre de résultats est limité à 1000, la taille du contexte à 300 caractères avant et après la chaîne de caractères recherchée et les « métadonnées » exportées sont l'auteur, le titre et la date de l'œuvre, à l'exception du sous-titre éventuel et de la pagination.

Le résultat de la recherche est envoyé dans le menu « téléchargement / download » de l'ordinateur. Nous verrons dans le chapitre 3 comment l'importer dans un tableur EXCEL.

1.2. *Intérêt et limites des recherches "simples"*

Il est nécessaire ici d'introduire les deux notions de « PRE-TRAITEMENT » et « POST-TRAITEMENT ».

- Tout traitement du corpus (sélection d'un type de recherche simple, assistée ou avancée, d'un mode de consultation et d'export des résultats) qui est effectué à l'aide des outils fournis par FRANTEXT-2 constitue un pré-traitement.

- Une fois que le résultat de la recherche est exporté dans un tableur, les opérations effectuées (tris, filtrages, ouverture de champs de commentaires, calculs de fréquence absolue et relative, etc.) relèvent du post-traitement.

En conséquence, tout pré-traitement détaillé est destiné à alléger le post-traitement et inversement tout pré-traitement grossier est destiné à déboucher sur un post-traitement détaillé.

Rechercher par la modalité « simple » d'une part toutes les occurrences de la forme « goût » (121) et d'autre part toutes celles de la forme « goûts » (9) dans le corpus des 4 romans d'Aragon présente l'intérêt de mettre en évidence la disparité entre les nombreuses occurrences au singulier et les rares occurrences au pluriel. Mais la procédure présente l'inconvénient d'entraîner deux exports distincts qu'il s'agit ensuite de faire figurer dans un même classeur EXCEL et de préférence dans une même feuille (► chap.3).

Le même résultat peut être obtenu plus simplement en effectuant une recherche « assistée » par LEMME (la forme privilégiée du mot qui figure comme tête des articles de dictionnaires, le singulier pour les noms, le masculin singulier pour les adjectifs, l'infinitif actif pour les verbes) englobant toutes les formes fléchies du mot. Une fois l'export du résultat disposé dans une feuille EXCEL, il suffit de trier alphabétiquement les lignes du champ « pivot » pour distinguer les lignes contenant en pivot soit *goût*, soit *goûts*. On peut également filtrer dans cette colonne l'une des deux formes du mot, ce qui permet de ne faire figurer que les unes ou les autres lignes et de faire apparaître le nombre d'occurrences concernées (► chap.3).

⇨ Le chapitre 2 sera consacré aux deux autres modes de recherche, ASSISTÉ et AVANCÉ.