La BFM 2022

Un corpus pour les recherches diachroniques en français médiéval et au-delà

ALEXEI LAVRENTIEV &
CÉLINE GUILLOT-BARBANCE

IHRIM / ENS DE LYON / CNRS

Plan

- Introduction : 33 ans d'histoire
- Représentativité
  - le corpus BFM2022
  - équilibrage du corpus
  - exploitation avec TXM 0.8.2
- Standardisation et interopérabilité
  - typologie textuelle
  - étiquettes UD
- BFM et corpus diachroniques longs
  - Passage du latin au français
  - Grande grammaire historique du français
  - intégration avec les Bibliothèques Virtuelles Humanistes
- Perspectives

La BFM : 33 ans d'histoire

- **1989** : Début du projet (C. Marchello-Nizia)
  complément au DMF et au Frantext (INALF / ATILF) : même choix méthodologique
- **1999** : Projet *Queste du Graal*, édition du ms. Lyon BM PA 77 (texte, traduction, images ms)
- **2000** : Base accessible en ligne, Weblex
- **2002 – 2005** : Relectures et encodage XML-TEI, échange avec l'ATILF
- **2005** : Ouverture du site du projet BFM http://bfm.ens-lyon.fr
- **2007- 2010** : Fermeture de la Base, contentieux avec la librairie Droz
- **2012** : Portail BFM http://txm.bfm-corpus.org, plateforme TXM
  BFM2012 (3,3 millions de mots), BFM2013 (4 700 000 mots), BFM2014 (3 550 000 mots), BFM2016 (4 100 000 mots),
- BFM2019 (4 700 000 mots)
- **BFM2022 (6 400 000 mots)**

BFM2022 : Nouveautés

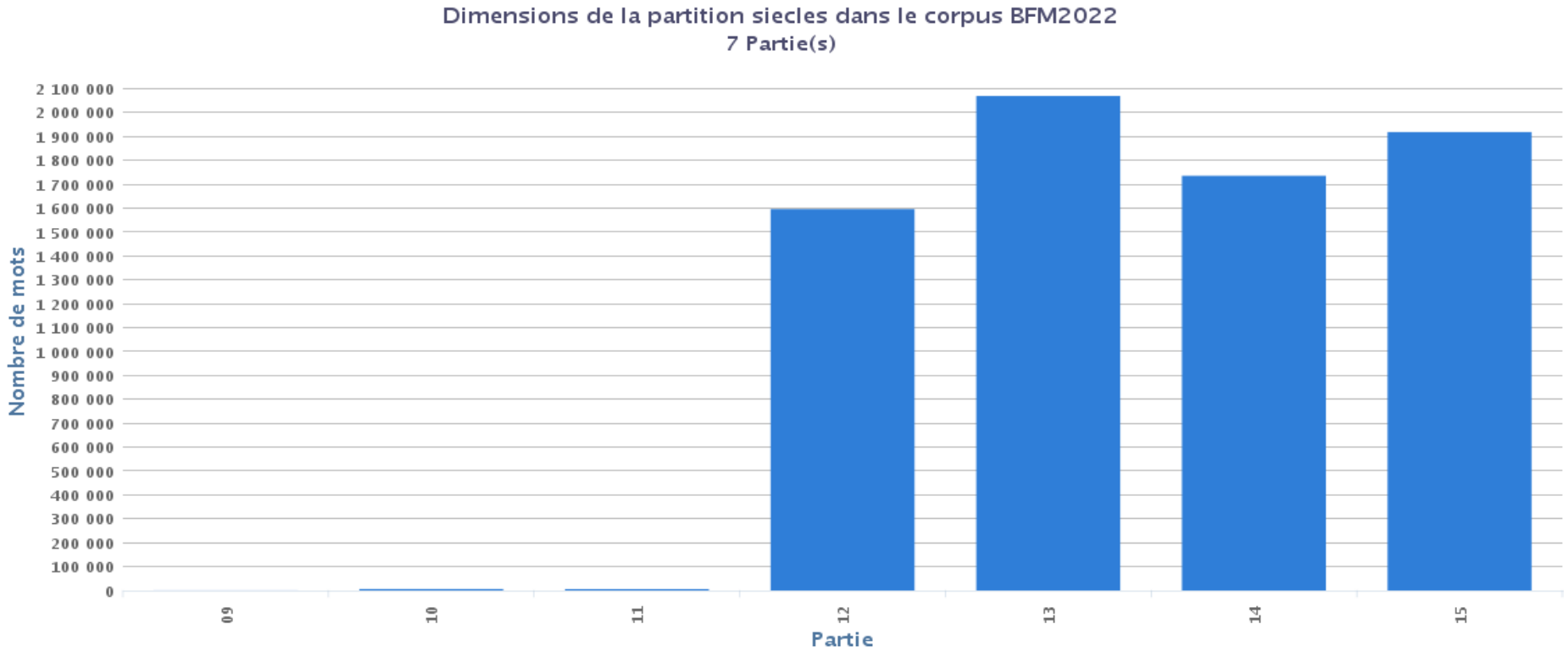
- 49 nouveaux textes
  - 170 → 219 textes
  - 4,7 → 6,4 millions de mots (+ 36%)
- Lemmatisation et étiquetage vérifiés
  - ajout de 11 textes (356 K → 612 K mots) lemmatisés
  - ajout de 8 textes (800 K → 1 M mots) étiquetés
- Nouveau corpus PROFITEROLE-V1-0
  - annotation syntaxique
- Changement d'hébergeur (→ Huma-Num) et mise à jour logicielle
  - (presque) transparents pour les utilisateurs
    - l'ancien portail restera accessible jusqu'en juin 2023
  - plus d'avertissements sur la connexion non sécurisée
  - assistant de requête
- Publication novembre 2022
  - accessible en version de test dès à présent : https://txm-bfm.huma-num.fr/txm

BFM2022 en bref

- Accès ouvert et gratuit
  - lecture des textes (HTML et PDF) libre
  - accès au moteur de recherche sur simple inscription
  - accès libre aux fichiers XML-TEI (Nakala)
  - téléchargement de corpus « binaire » pour TXM 0.8.2
- Sous-corpus « prêts à l'emploi »
  - étiqueté (vérifié)
  - lemmatisé (vérifié)
  - CORPTEF (corpus indépendant dans l'ancien portail)
  - périodisation par siècle (9-11, 12, 13, 14, 15)
- Corpus supplémentaires
  - PROFITEROLE, PALAFRA, BFM-MSS
  - BFM2019 (jusqu'à la fin de l'année 2023)
- Éditions originales
  - *Queste del saint Graal*
  - *St Alexis, Psautier d'Arundel, Quinze Joyes, Image du monde* (prototypes)

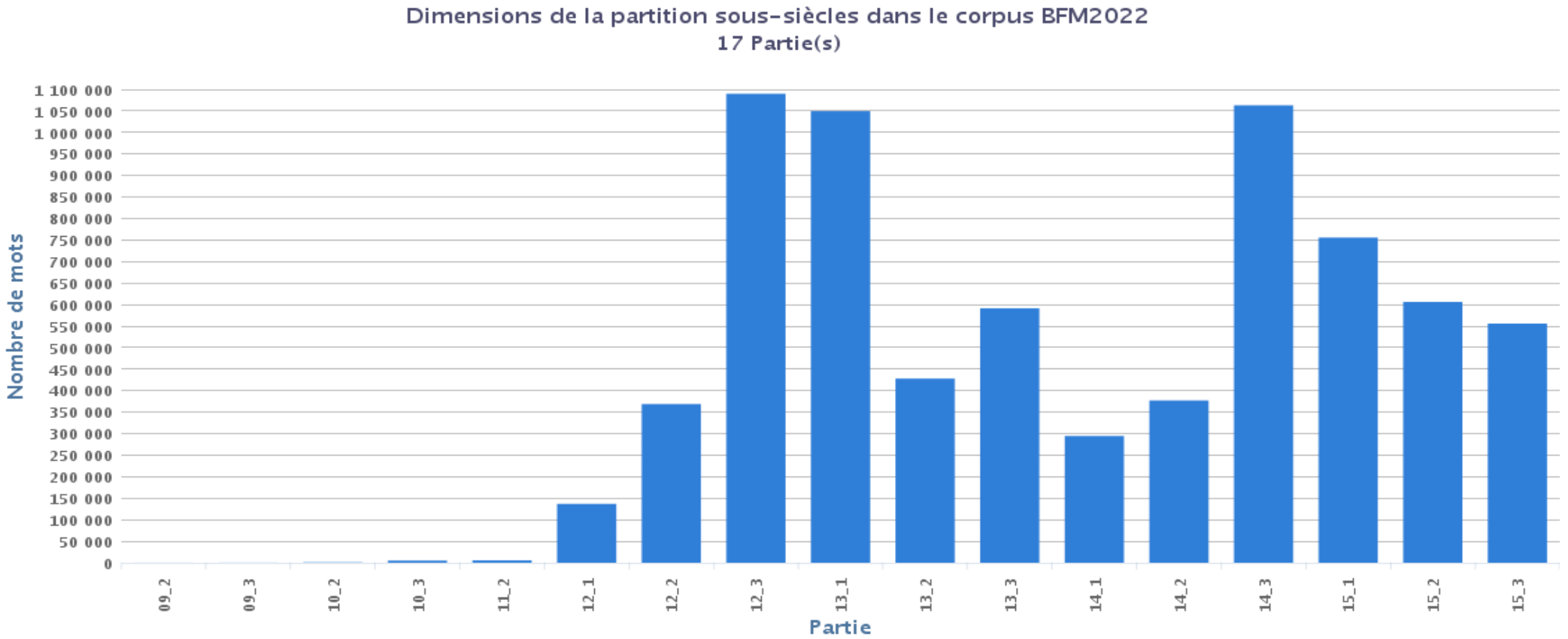
BFM2022 : Représentativité

- Siècles
  - entre 1,6 et 2,0 M mots du 12e au 15e



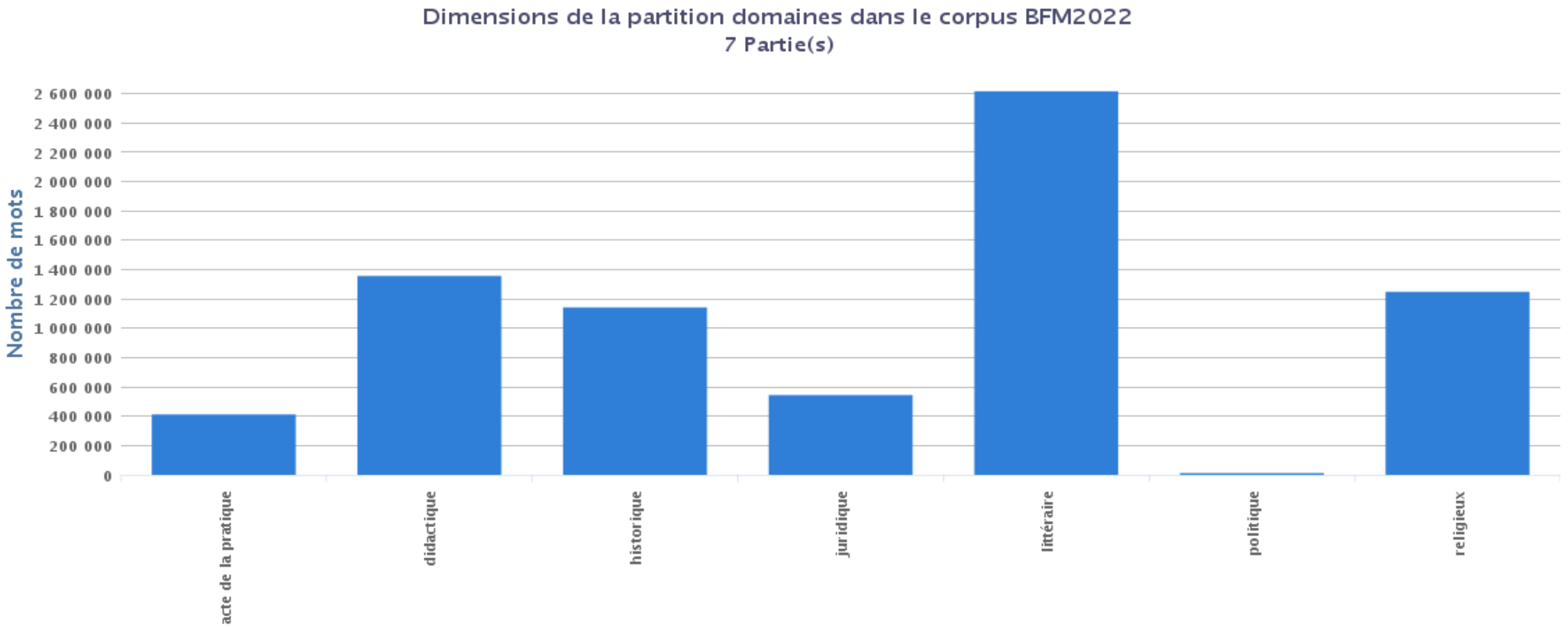
BFM2022 : Représentativité

- Sous-siècles (début / milieu / fin)



BFM2022 : Représentativité

- Domaines

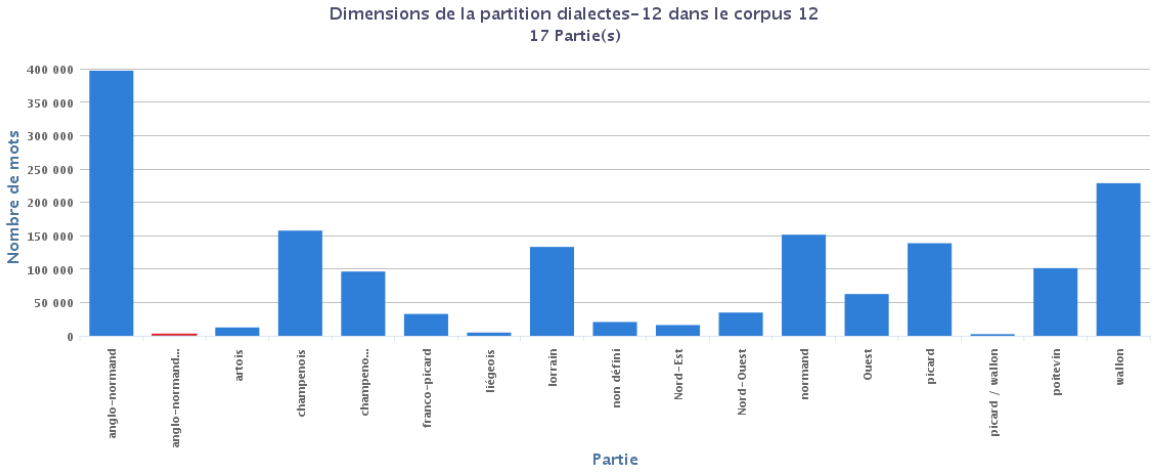


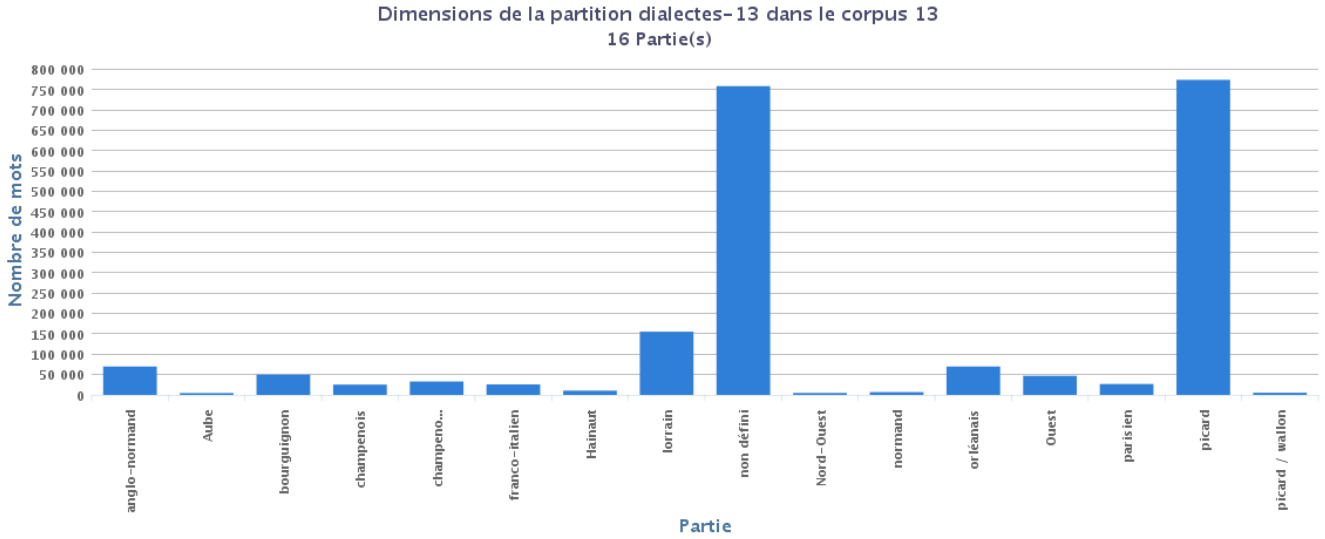
BFM2022 : Représentativité

- Domaines par siècle

BFM2022 : Représentativité

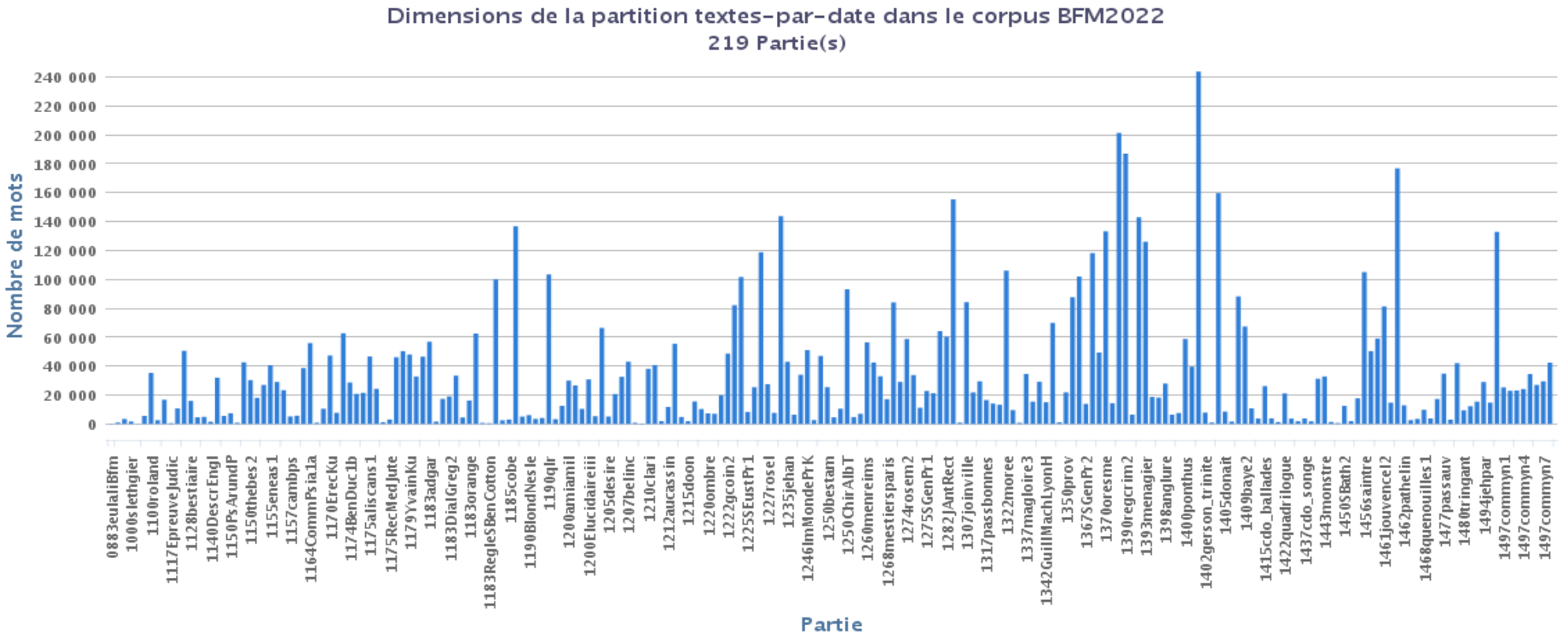
- Dialectes (12 - 13)





BFM2022 : Représentativité

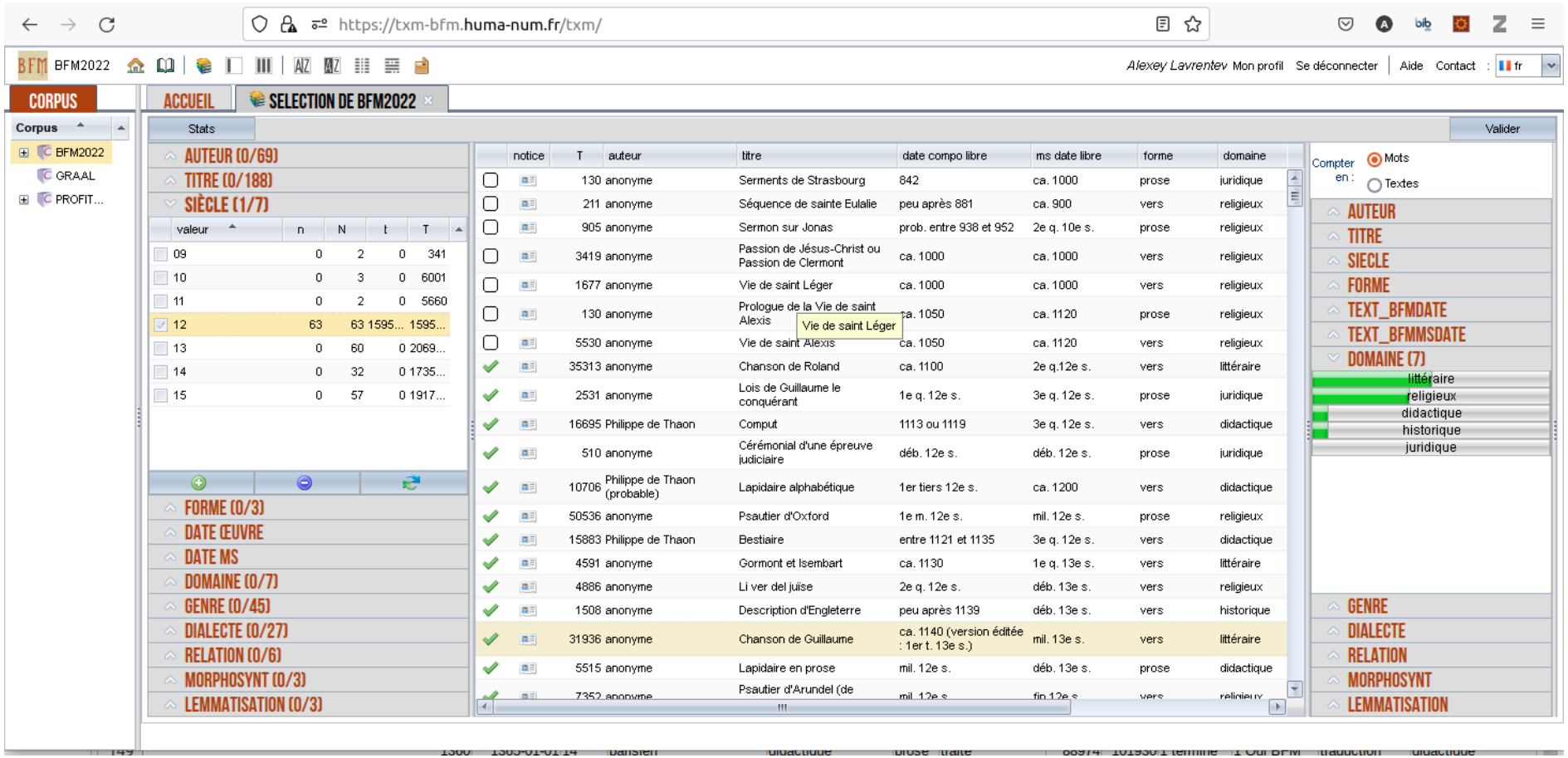
- Taille des textes



BFM2022 : Représentativité

- Synthèse
  - faire attention à la distribution par partie
    - e.g. dialectes par siècle
  - faire attention à de très gros textes
    - e.g. Registre criminel du Châtelet : 350 000 mots

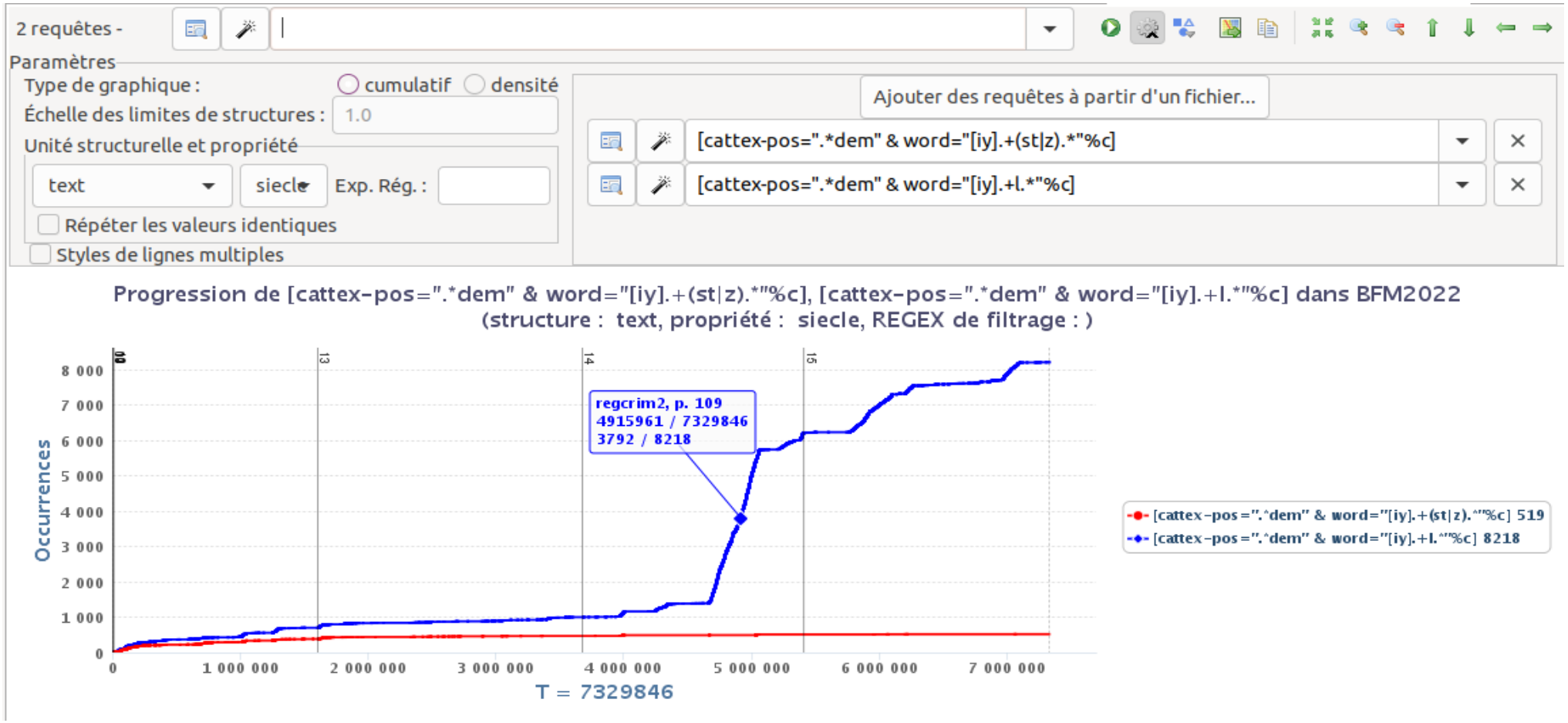
BFM2022 : Sélection de textes



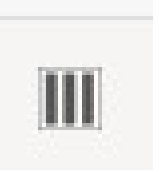
BFM2022 avec TXM 0.8.2

- Télécharger et installer TXM
  - http://textometrie.org → https://txm.gitpages.huma-num.fr/textometrie/files/software/TXM/0.8.2/
    - attention : pour le moment TXM n'est pas compatible avec les Macs équipés d'un processeur M1
- Télécharger le corpus « bfm2022.txm »
  - https://txm-bfm.huma-num.fr/txm
  - sélectionner le corpus BFM2022 et cliquer sur 
- Ouvrir TXM, puis utiliser la commande « Fichier > Charger > un corpus binaire (.txm)... »

BFM2022 avec TXM 0.8.2 : Progression

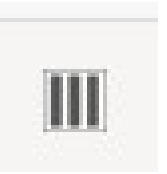


BFM2022 avec TXM 0.8.2 : Partition



- Mode « Simple »
  - sélectionner l'unité de découpage (« text » le plus souvent) et la propriété (métadonnée) de partition (siècle, domaine, dialecte…)
- Mode « Assisté »
  - sélectionner à la main les unités à mettre dans une partie
  - attention à éviter le recouvrement et des trous entre les parties
- Mode « Avancé »
  - générer chaque partie avec une requête CQL
  - attention au risque d'atomisation (les requêtes sur des syntagmes peuvent ne plus être possibles)
- → Visualiser les dimensions des parties (commande « Dimensions »)

BFM2022 avec TXM 0.8.2 : Partition



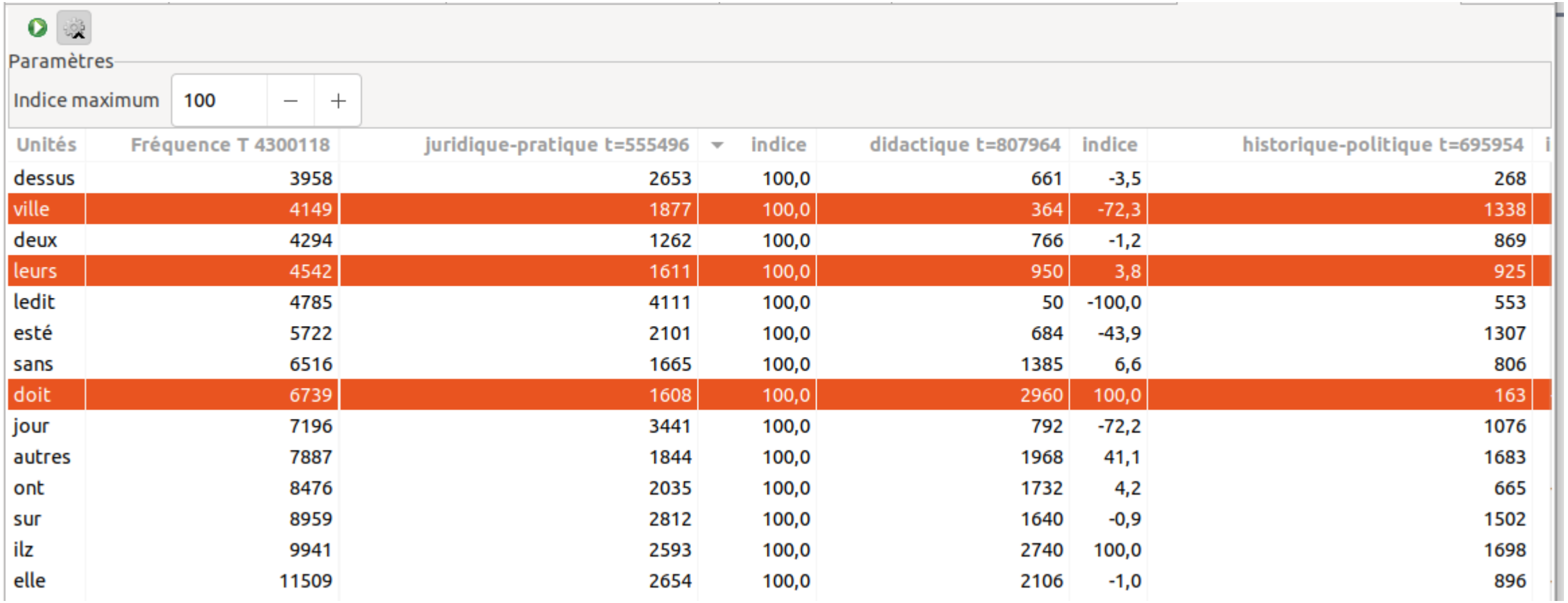
- Pour équilibrer une partition
  - utiliser le mode « Assisté »
  - ou créer d'abord un sous-corpus sans les textes sur-représentés
  - Il n'y a pas d'interface de sélection dynamique de textes comme sur le portail
- Pour échantillonner des textes
  - créer un sous-corpus en mode avancé en sélectionnant des divisions à garder (pas pratique)
  - récupérer et éditer des fichiers XML TEI TXM, puis réimporter le corpus
    - Cette possibilité existe, mais pour le moment elle n'est pas documentée
    - Une connaissance du format XML et la maîtrise de logiciels spécialisés (type Oxygen XML) sont nécessaires

BFM2022 avec TXM 0.8.2 : Table lexicale

- Commande accessible sur une partition ou un index de partition
- Génère un tableau de formes (lemmes, étiquettes, etc.) avec leur fréquence par partie
- Ce tableau est éditable, on peut
  - supprimer ou fusionner des colonnes (parties)
  - supprimer ou fusionner des lignes (lemmatiser, annoter…)
  - l'enregistrer pour la réutilisation

- Les calculs statistiques (AFC, CAH, Spécificités) se font toujours sur une table lexicale… même si on peut « sauter » l'étape de son édition

BFM2022 avec TXM 0.8.2 : Spécificités

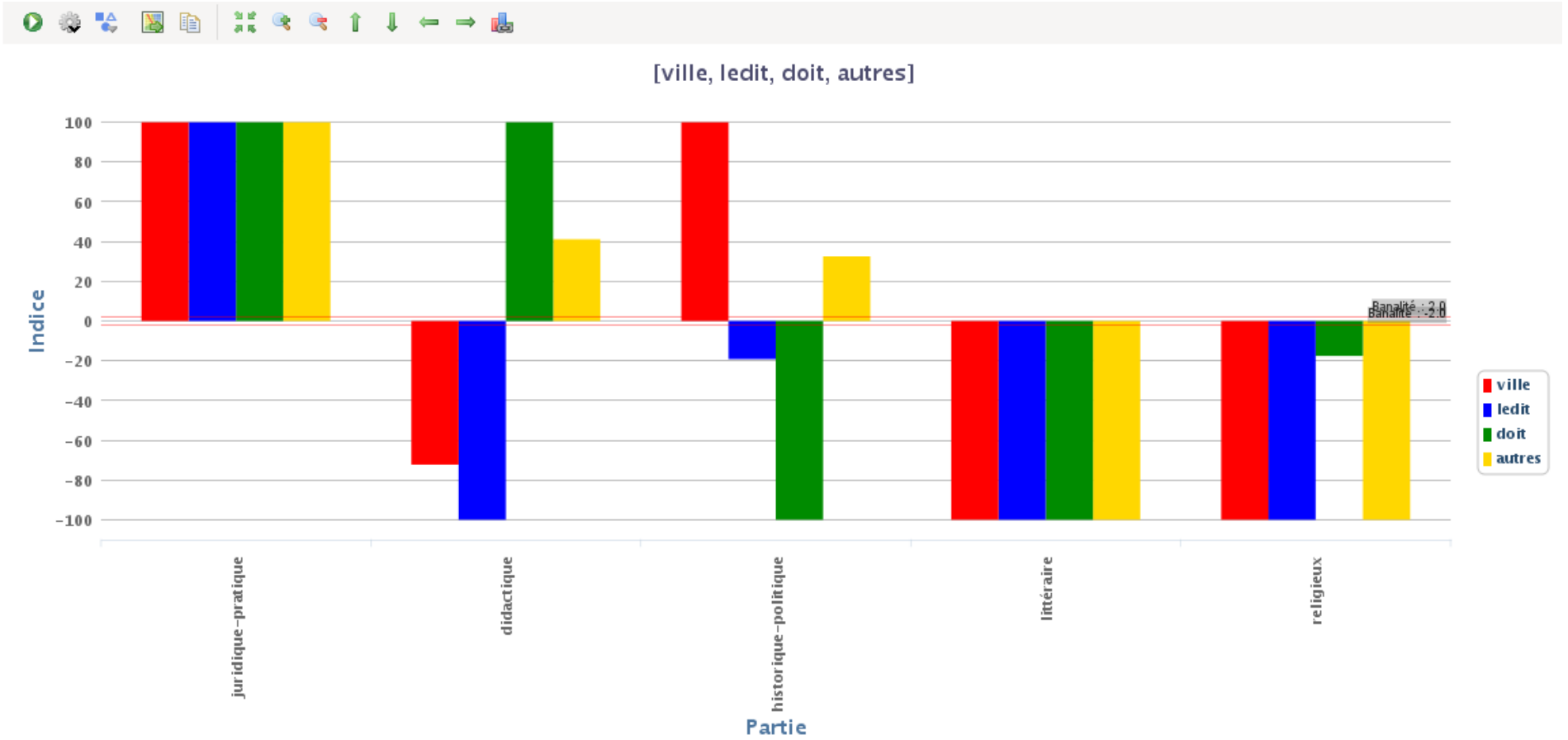


Paramètres

Indice maximum 100

Unités	Fréquence T 4300118	juridique-pratique t=555496	indice	didactique t=807964	indice	historique-politique t=695954	i
dessus	3958	2653	100,0	661	-3,5	268	
ville	4149	1877	100,0	364	-72,3	1338	
deux	4294	1262	100,0	766	-1,2	869	
leurs	4542	1611	100,0	950	3,8	925	
ledit	4785	4111	100,0	50	-100,0	553	
esté	5722	2101	100,0	684	-43,9	1307	
sans	6516	1665	100,0	1385	6,6	806	
doit	6739	1608	100,0	2960	100,0	163	
jour	7196	3441	100,0	792	-72,2	1076	
autres	7887	1844	100,0	1968	41,1	1683	
ont	8476	2035	100,0	1732	4,2	665	
sur	8959	2812	100,0	1640	-0,9	1502	
ilz	9941	2593	100,0	2740	100,0	1698	
elle	11509	2654	100,0	2106	-1,0	896	

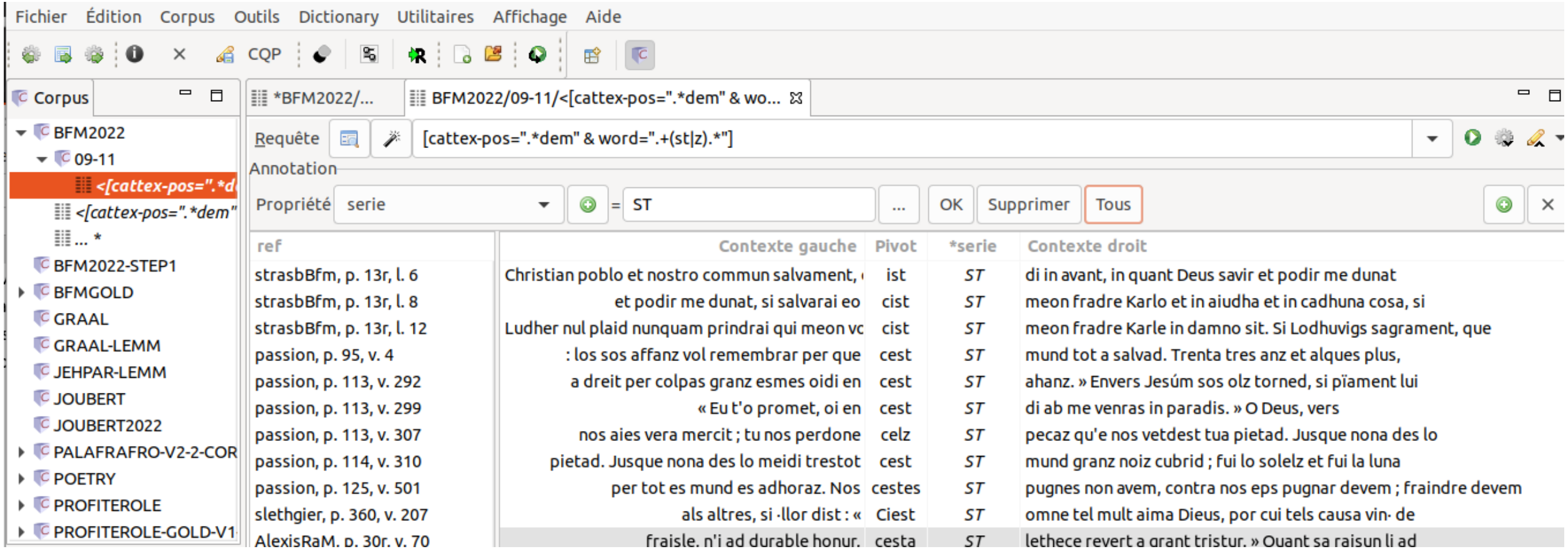
BFM2022 avec TXM 0.8.2 : Spécificités



TXM 0.8.2 : Annotation

- Plusieurs types d'annotation sont proposées
  - Unité-Relation-Schéma (cf. le projet DEMOCRAT)
  - Propriétés de mots en concordance
- La fonctionnalité n'est pas adaptée à de gros corpus
  - La sauvegarde d'annotations sur la BFM2022 risque de prendre plusieurs heures…
  - → privilégier des corpus de travail restreints
  - *Des paramètres de sauvegarde permettent d'optimiser et accélérer le processus*
- Voir les support de l'atelier BFM-TXM au colloque Diachro X
  - https://halshs.archives-ouvertes.fr/halshs-03788509

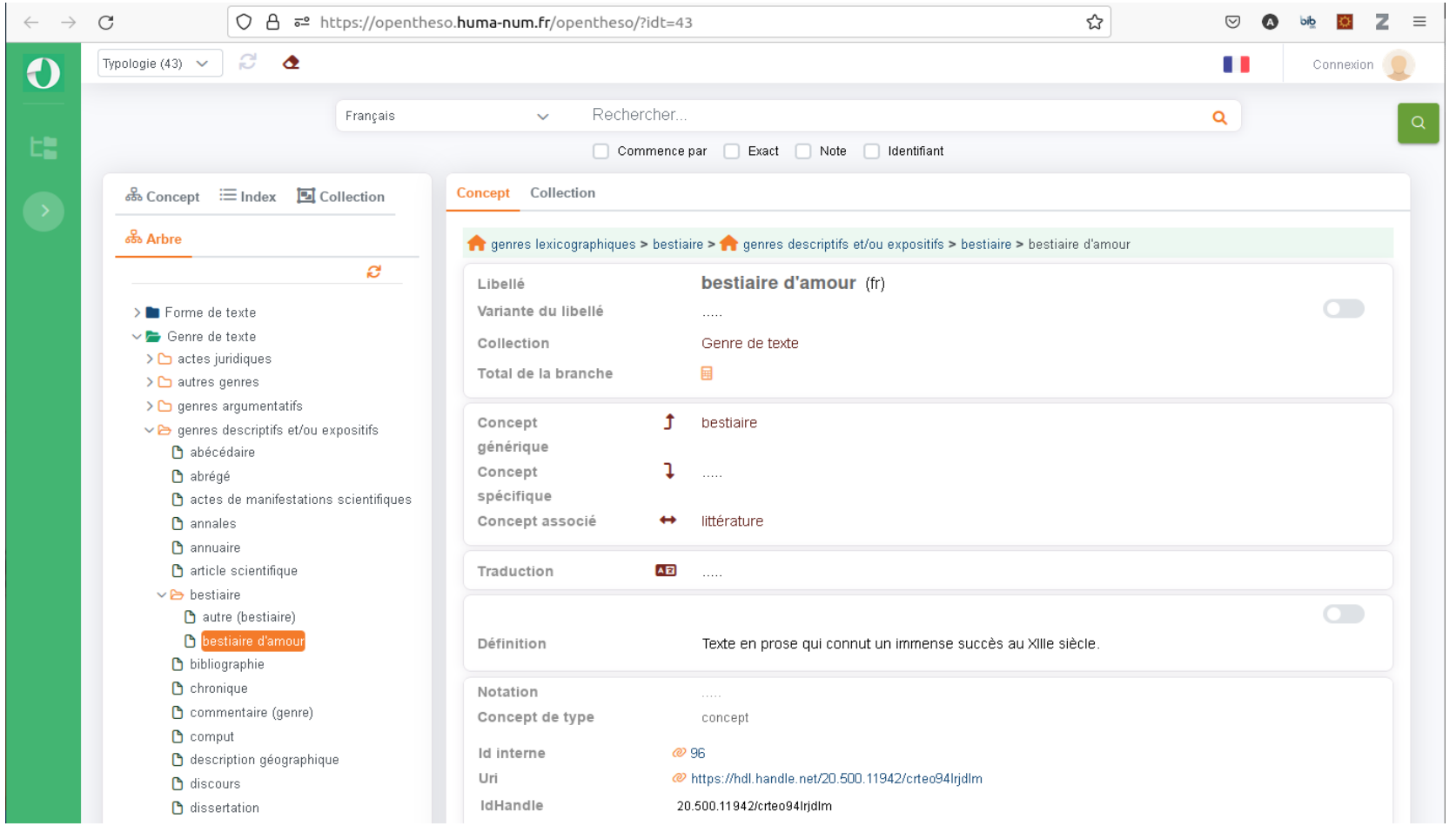
TXM 0.8.2 : Annotation



Standardisation et interopérabilité

- Thesaurus « Typologie » (consortium CAHIER)
  - https://opentheso.huma-num.fr/opentheso/api/theso/Typologie
  - 366 « concepts » permettant de décrire et classer des textes
  - 9 « collections » = points de vue sur les textes
    - Domaine, Factualité, Forme, Genre, Mode d'agencement, Origine, Type de discours, Public cible, Mode d'inscription
  - + 3 branches techniques
    - Classe de descripteur, Chronologie, Langue du texte
  - Tous les termes sont définis, organisés hiérarchiquement et dotés d'un identifiant pérenne (Handle)

Thésaurus « Typologie textuelle »



Standardisation et interopérabilité

- Étiquettes Universal Dependencies https://universaldependencies.org
  - ud-pos : parties du discours
    - NB : la BFM ne distingue pas les verbes principaux et auxiliaires (sauf dans le corpus PROFITEROLE)
  - ud-feats : sous-catégories ou traits morphologiques
    - NB : seuls les traits enregistrés dans Cattex-min sont indiqués
      - PronType=Dem, VerbForm=Fin
      - mais pas Number=... ou Tense=...
  - les contractions (*del, sis*) portent une double étiquette

Standardisation et interopérabilité

- **Licences**
- Depuis le procès Droz contre Classiques Garnier (2014, appel 2017) le texte « brut » des éditions de textes médiévaux est considéré comme étant dans le domaine public.
  - aucune restriction pour l'affichage du texte et pour la taille des contextes
- ⇒ Licence ouverte Etalab
  - pour le corps du texte
  - pour les suppléments numériques (balisage TEI, annotations)
- ⇒ Licence Creative Commons BY-NC-SA
  - pour l'apparat critique (notes, introduction...)

- Les fichiers XML-TEI sont mis à disposition dans l'entrepôt NAKALA
  - https://nakala.fr/u/collections/10.34847/nkl.1279lie9

BFM et corpus diachroniques « longs »

- Exemples d'extensions diachroniques
  - Grand Frantext (2002-2005)
    - Echange d'une 40aine de textes
    - Intégration dans les deux bases (Frantext/BFM)
  - Grande grammaire historique du français (2007-2020)
    - Projet éditorial de grande ampleur
    - Corpus original à la base de la Grammaire
  - Passage du latin au français (2015-2018)
    - Projet ANR (ENSL et U. de Lille) / DFG (U. de Regensburg et Tübingen)
    - Corpus bilingue latin-français
  - Corpus DEMOCRAT (2016-2020)
    - Projet ANR (Lattice, ENSL et U. de Strasbourg)
    - Annotation de chaînes de référence
  - Corpus « démonstratifs »
    - Corpus adhoc pour une recherche en cours (2022)
    - Intégration Moyen Âge - 16e siècle

BFM et corpus diachroniques « longs » : remontée vers le latin mérovingien

Passage du latin au français (PaLaFra)

- Sous-corpus / Sources
  - corpus latin (6e-8e s. , 350 000 mots) / Monumenta Germaniae Historica
  - corpus français (9-13e s., 1 M de mots) / BFM
  - corpus aligné (4 textes, 83 000 mots) / BFM
- Critères de sélection pragmatiques (métadonnées BFM)
  - « scripta latina rustica » (Sabatini 1968)
  - études longitudinales latin-français : vies de saints, chartes, textes historiques
- Étiquetage morphosyntaxique (Cattex2009 / lapos / UD)
- Lemmatisation (Latin-German lexicon / DMF)
- Diffusion sur le Portail BFM et sous forme de corpus “binaire” TXM (licences)

BFM et corpus diachroniques « longs » : descente jusqu'au français contemporain

Grande grammaire historique du français (9e - 20e s.)

- Sources
  - BFM
  - Bibliothèques virtuelles humanistes
  - DMF et Frantext
- Métadonnées unifiées
  - auteur/ titre / date / siècle / forme / domaine / genre / dialecte
- Corpus noyau (échantillonné et équilibré) / corpus intégral
- Étiquetage morphosyntaxique
  - modèle linguistique pour le français médiéval et le 16e s. (cattex2009)
  - modèle linguistique pour le français du 17e au 20e s. (French TreeTagger)
- Diffusion sous forme de corpus « binaire » TXM

BFM et corpus diachroniques « longs » : intégration Moyen Âge- 16e siècle

Corpus « démonstratifs »

- ensemble des corpus BFM2022 et BVH-EPISTEMON + les textes du 16e s. de la GGHF
- lemmatisation automatique des démonstratifs (par une macro TXM)
- annotation plus fine dans une sélection de textes

Perspectives

- Outils pour la construction de corpus étendus
  - tutoriels spécialisés
  - formation des doctorant(e)s et enseignants-chercheurs
  - création d'un « kit » de corpus diachronique pour TXM
  - Open French Corpus (CORLI)
- Noces de philologie, Linguistique et Numérique
  - section "Philologie linguistique et corpus médiévaux"
  - éditions/Corpus : projet Fabliaux
- Extension du corpus au début 16e siècle
- Développement de corpus multilingues
  - occitan
  - autres langues romanes