BDLC : Accès aux ethnotextes par concordances

Laurent Kevers

BDLC : Accès aux ethnotextes par concordances

Guide d'utilisation

Laurent Kevers
UMR CNRS 6240 LISA
Université de Corse Pascal Paoli
kevers_l@univ-corse.fr
Octobre 2021

1 Introduction

1.1 Qu'est-ce qu'un concordancier…

Le concordancier est un outil qui permet, au travers de requêtes plus ou moins complexes, d'explorer un corpus de textes, c'est-à-dire « une collection de données langagières sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage » (Sinclair). Le résultat d'une requête est appelé une *concordance*. Celle-ci reprend les occurrences trouvées dans le corpus, accompagnées de leur contexte gauche et droit.

La linguistique de corpus permet, au travers d'outils tels que les concordanciers :

- d'étudier le langage ;
- à travers les productions authentiques des locuteurs ;
- orales ou écrites ;
- et non en utilisant uniquement ses propres compétences de locuteur.

Cet outil est a été intégré à la BDLC (accessible à l'adresse https://bdlc.univ-corse.fr/concord/) à partir d'un développement réalisé par le Cental[1] (Centre de Traitement Automatique du Langage, Université de Louvain-la-Neuve).

1.2 Corpus

La BDLC met à disposition un corpus d'ethnotextes qui ont été récoltés depuis le début du projet. Il s'agit de productions authentiques, récoltées sur le terrain. Ces textes sont en rapport avec les différents thèmes étudiés par la BDLC (culture, traditions et savoir faire en Corse). L'ensemble de documents accessible par l'intermédiaire de ce concordancier n'est pas figé et a vocation à être mis à jour régulièrement.

Le concordancier dispose actuellement d'une base de textes non lemmatisés. Cela implique qu'il est uniquement possible, pour l'instant, d'effectuer des recherches sur des formes brutes. A l'avenir, lorsque des textes lemmatisés et annotés morphosyntaxiquement seront disponibles, il sera possible de construire des requêtes en utilisant des éléments tels qu'un lemme ou une catégorie grammaticale.

1 https://uclouvain.be/fr/instituts-recherche/ilc/cental

2 Présentation générale de l'interface

L'interface se compose de trois zones, comme indiqué aux illustrations 1 et 2 :

1. Les filtres sur les métadonnées
2. Le champ d'entrée de la requête
3. La zone de visualisation des résultats (concordances)

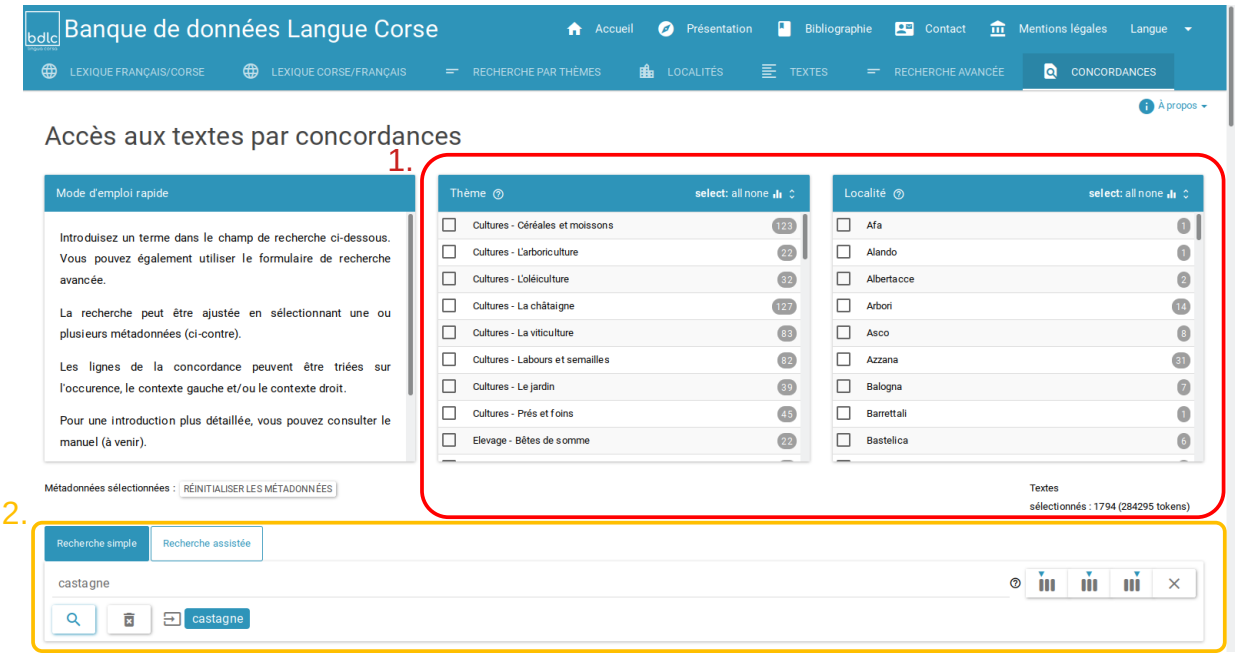

Illustration 1: Zones de filtrage sur les métadonnées (1) et d'introduction des requêtes (2)

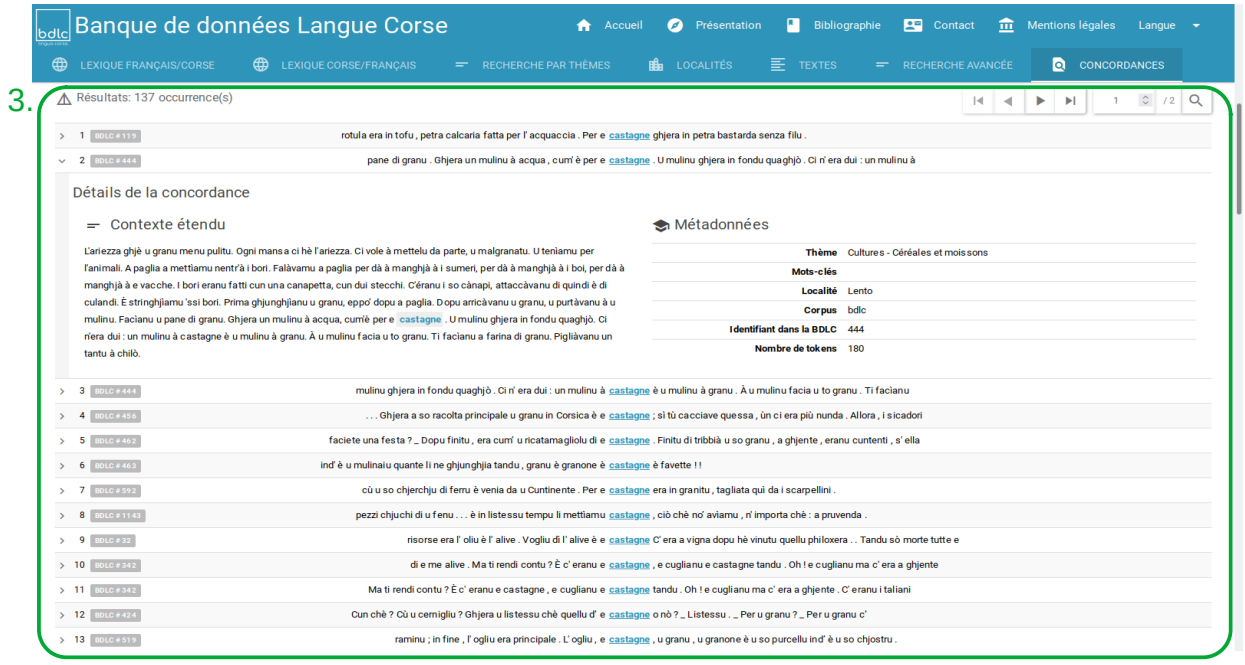

Illustration 2: Liste de résultats (3) sous forme de concordance, avec détails pour la deuxième occurrence

La liste de résultat apparaît, à partir du moment où une requête a été formulée, sous les zones dédiés au filtrage par métadonnées et à la recherche. Ces deux zones restent accessibles à tout moment afin de pouvoir modifier ou reformuler la requête.

La zone de recherche se présente par défaut sous la forme d'un formulaire de recherche simple (illustration 1 - point 2). Il est également possible d'utiliser un formulaire de recherche assistée, comme montré à l'illustration 3.

Illustration 3: La zone de requête (illustration 1 - point 2) dans sa version « assistée »

2.1 Les filtres sur les métadonnées

Actuellement, deux métadonnées sont exploitables. Il s'agit du thème de l'ethnotexte, ainsi que de la localité où il a été récolté.

Le nombre de textes correspondant à chaque métadonnée est indiqué en regard de celle-ci. Lorsqu'une ou plusieurs valeurs sont sélectionnées pour une métadonnée (par exemple le thème), le dénombrement est automatiquement actualisé pour l'autre métadonnée (par exemple la localité) afin de tenir compte de ce premier filtrage.

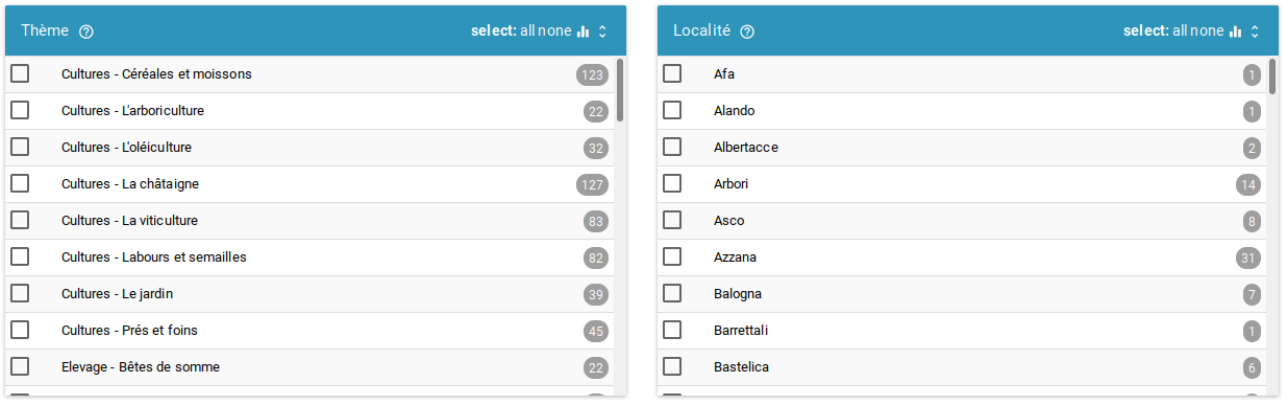

Illustration 4: Filtres sur les métadonnées

Lorsqu'un filtrage sur les métadonnées est mis en place, sa composition est rappelée grâce à l'apparition de petites étiquettes (voir illustration 5). Celles-ci permettent éventuellement de retirer un composant du filtre. Le filtre peut également être complètement annulé en cliquant sur le bouton « Réinitialiser les métadonnées ». Enfin, l'impact du filtre sur le nombre d'ethnotextes ciblés, ainsi que sur le nombre de tokens que ceux-ci représentent est également renseigné.

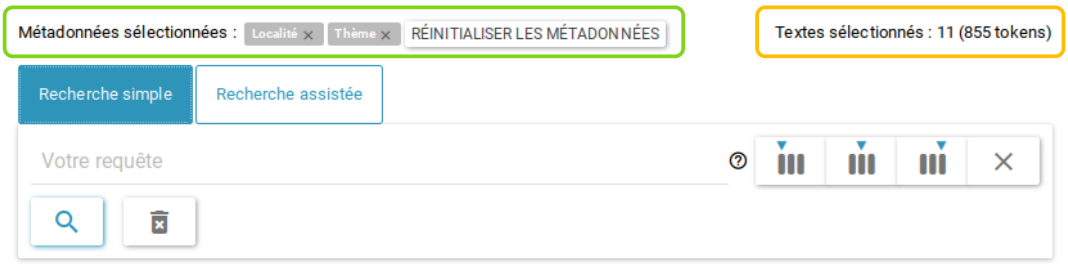

Illustration 5: Les filtres activés, ainsi que le nombre de textes et de tokens concernés, sont indiqués au dessus de la zone de requête.

2.2 La formulation d'une requête

La recherche peut être exprimée soit au moyen d'un champ de recherche classique (« recherche simple »), soit par l'intermédiaire d'un formulaire plus structuré qui permet d'accompagner l'utilisateur dans la construction de la requête (« recherche assistée»).

Pour la **recherche assistée** (illustration 3), l'expression de la requête passe par l'utilisation d'une ou plusieurs « fiches ». Une fiche est composée d'un formulaire qui reprend un ensemble de champs. Ceux-ci permettent de caractériser un terme de la requête correspondant à un mot ou token à rechercher. Il est possible d'éditer successivement plusieurs fiches afin de formuler une requête composée par un ou plusieurs mots/tokens. L'ajout d'une fiche, c'est-à-dire d'un terme, à la requête peut être effectué en cliquant sur l'icône « + » située au dessus des fiches.

Pour un terme, ou une fiche, il est nécessaire de remplir une des trois lignes proposées :

1. forme, sans que la présence de majuscules/minuscules soit prise en compte ;
2. forme, avec prise en compte des majuscules/minuscules ;
3. n'importe quel mot (case à cocher).

Pour les deux premier cas, il est nécessaire d'introduire une chaîne de caractères dans le champ texte correspondant. Une liste déroulante permet également de spécifier si cette chaîne correspond à un mot entier (option « égal ») ou à une portion de mot (options « commence par », « se termine par », « contient »). On peut également choisir l'option « différent » qui correspondra à n'importe quel mot, sauf celui spécifié. Enfin l'option « ou » permet d'exprimer une disjonction de plusieurs termes, ceux-ci devant alors être séparés par un espace lors de leur introduction dans le champ texte.

Il est à noter que les champs permettant l'introduction des chaînes de caractère proposent une fonctionnalité d'aide à l'introduction (« auto-complétion ») qui propose, dès l'introduction des premiers caractères, une liste de mots présents dans le corpus. Il n'est cependant pas obligatoire de sélectionner un élément de cette liste, on peut introduire la chaîne de caractères de son choix.

Une fois la fiche remplie, son contenu est traduit dans la syntaxe propre au moteur de recherche du concordancier, et est affiché dans un onglet au dessus du formulaire (illustration 6). Il est alors

possible de continuer l'expression de la requête, si celle-ci contient plus qu'un mot, en cliquant sur l'icône « + » qui permet d'ouvrir un nouveau formulaire vierge.

Si l'ordre des différents mots d'une requête doivent être réordonnées, il est possible de déplacer chaque élément en utilisant les boutons « déplacer à gauche » ou « déplacer à droite » présents au bas de chaque fiche. Il est bien entendu également possible de supprimer un élément, voir l'ensemble de la requête.

La requête finale est reprises en bas du cadre de recherche, juste à côté du bouton permettant de lancer la recherche, ainsi que de celui destiné à réinitialiser l'ensemble de la requête (illustration 6).

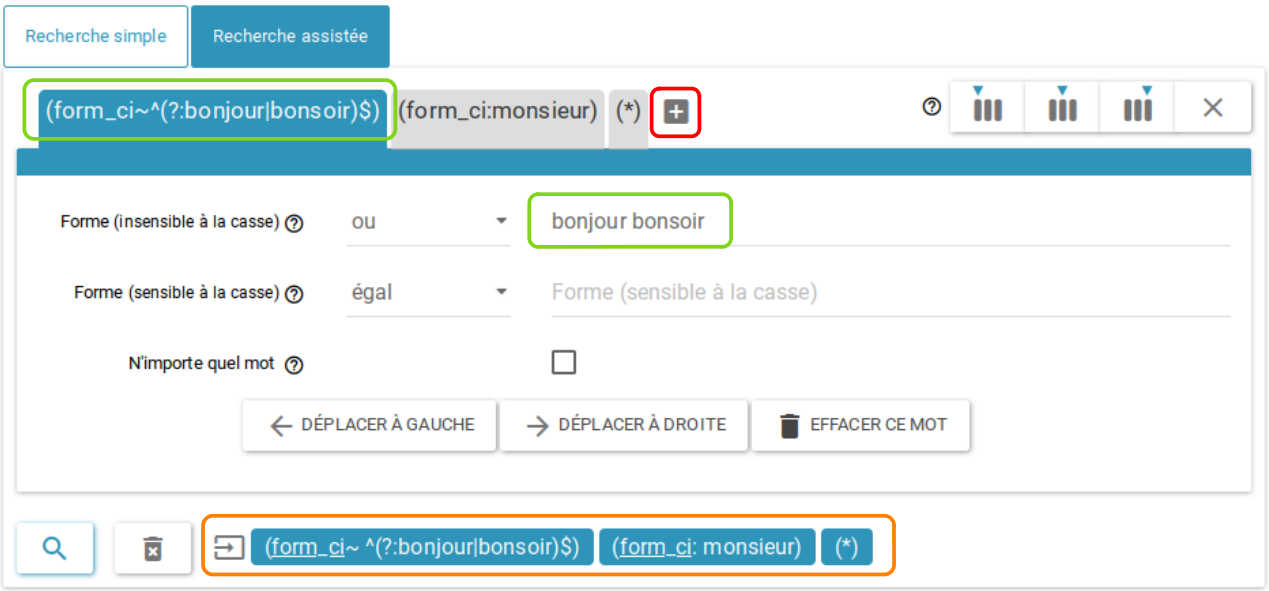

Illustration 6: Le texte "bonjour bonsoir" introduit dans le formulaire est traduit – en fonction des options choisies et selon le formalisme interne du concordancier – pour former un terme de la requête. Celui-ci est visible sur l'onglet de la fiche. La requête complète, formée par l'ensemble des fiches est reprise en bas.

Enfin, les options de tri permettent de choisir dans quel ordre seront présentés les résultats. Par défaut, aucun tri n'est appliqué. Les occurrences issues d'un même texte apparaissent cependant les unes à la suite des autres. L'ordre alphabétique peut être imposé zone par zone – occurrence, contexte gauche, contexte droit – en décidant de l'ordre d'application du tri. Par exemple, si l'on désire obtenir un tri sur l'occurrence, puis sur le contexte gauche, il faudra cliquer sur l'icône du milieu et ensuite sur celle de gauche. Des étiquettes numérotées apparaissent alors afin de matérialiser le tri demandé (illustration 7). Après chaque modification des options de tri, il est nécessaire de relancer la recherche afin que les nouveaux paramètres soient pris en compte.

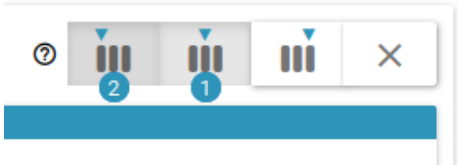

Illustration 7: Activation du tri

Le mode de **recherche simple** se résume à l'utilisation d'un unique champ de texte. Celui-ci permet de formuler des recherche très simples telles que par exemple la forme « cane ». La requête peut cependant également être introduite selon la syntaxe propre au moteur de recherche du concordancier. Lorsque cette syntaxe est maîtrisée, ce mode de recherche permet un introduction plus rapide de la requête, tout en offrant diverses possibilités qui rendent ce mode de recherche plus complet et puissant.

Le tableau ci-dessous reprend une liste, accompagnée d'exemples, des principaux éléments du langage de requête. Ces requêtes couvrent les possibilités offertes par la recherche assistée.

Requête	**Signification**	**Exemple de résultat**
`(form:abcde)`	Forme égale à 'abcde', sensible à la casse	*abcde*
`(form_ci:abcde)`	Forme égale à 'abcde', insensible à la casse	*abcde, Abcde ...*
`abcde`	Équivalent à `(form_ci:abcde)`	
`(form!:abcde)`	Forme différente de 'abcde', sensible à la casse	*xyz, uvw, Abcde ...*
`(form_ci!:abcde)`	Forme différente de 'abcde', insensible à la casse	*xyz, uvw ...*
`(form~^abcde)`	Forme commençant par 'abcde', sensible à la casse	*abcdefgh ...*
`(form_ci~^abcde)`	Forme commençant par 'abcde', insensible à la casse	*abcdefgh, Abcdefgh ...*
`(form~xyz$)`	Forme se terminant par 'xyz', sensible à la casse	*uvwxyz, Uvwxyz ...*
`(form_ci~xyz$)`	Forme se terminant par 'xyz', *in*sensible à la casse	*uvwxyz, Uvwxyz, uvwXyz ...*
`(form~bcd)`	Forme contenant 'bcd', sensible à la casse	*bcd, abcd, bdce, abcde ...*
`(form-ci~bcd)`	Forme contenant 'bcd', insensible à la casse	*bcd, abcd, bdce, abcde* *Bcd, abCd, bcDe, aBCDe …*
`(form~^(?:abc|xyz)$)`	Forme égale à 'abc' ou égale à 'xyz', sensible à la casse	*abc, xyz*
`(form_ci~^(?:abc|xyz)$)`	Forme égale à 'abc' ou égale à 'xyz', insensible à la casse	*abc, Abc, aBc, xyz, Xyz, xyZ…*
`(*)`	N'importe quel mot	*Abcde, xyz, 23 ...*

L'utilisation de la recherche simple permet en réalité d'exploiter toutes les possibilités offertes par l'implémentation des expressions régulières POSIX dans PostgreSQL. La documentation à ce sujet peut être consultée à l'adresse suivante :

https://www.postgresql.org/docs/13/functions-matching.html#FUNCTIONS-POSIX-REGEXP.

Ci-dessous, quelques exemples de requêtes allant au-delà des possibilités offertes par la recherche assistée. Cette liste n'est pas exhaustive, mais reprend les éléments les plus courants.

- Utilisation du joker « . » : ce caractère spécial remplace n'importe quel caractère :

  ```
  (form_ci~^castagn.$)
  ```

- Utilisation d'une classe de caractères « [abcde] » : cette expression permet de reconnaître n'importe quel caractère énuméré entre les crochets

  ```
  (form_ci~^castagn[ei]$)
  ```

- Certaines classes de caractères peuvent être définies par des intervalles, « [x-z] » :

  `(form_ci~^[a-z]$)` : lettres minuscules non accentuées

  `(form_ci~^[0-9]$)` : chiffres de 0 à 9

- Il est possible d'appliquer une négation sur une classe de caractères, « [^abcde] » :

  ```
  (form_ci~^castagn[^ei]$)
  ```

- La quantification (répétition) 0-N peut être exprimée par le caractère spécial « * » :

  `form_ci~^1[0-9]*$)` : succession de chiffres quelconque d'une longueur indéterminée commençant par '1'

- La quantification (répétition) 1-N peut être exprimée par le caractère spécial « + » :

  `form_ci~^[0-9]+$)` : succession de chiffres d'une longueur minimale de 1

- D'autres modes de quantification sont possibles :
  - {m} : exactement m éléments :

    ```
    form_ci~^[0-9]{4}$)
    form_ci~d{2}.*r{2}.*z{2})
    ```

- Les modes ci-dessous sont autorisés par la norme POSIX, **mais ne sont pas supportés** dans le concordancier :
  - {m,} : minimum m éléments :

    ```
    (form_ci~[aeiou]{3,})
    ```

  - {m,n} : entre m et n éléments :

    ```
    (form_ci~[aeiou]{3,5})
    ```

- Il n'est également **pas possible** de spécifier la virgule « , » en tant que token à rechercher.

La recherche simple permet, comme dans le cas de la recherche assistée d'effectuer des requêtes composées de plusieurs termes :

```
(form:castagne)(*)
(form:castagne)(form:bianche)
(form_ci~^(castagn|teghj))(form_ci~^[a-z])
```

2.3 La zone de visualisation des résultats (concordances)

Cette zone s'affiche, lorsque des résultats sont disponibles, sous la zone de recherche (sur l'illustration 2, la zone de recherche n'apparaît pas suite au défilement de la page dans le navigateur).

Le nombre de résultats est renseigné dans la partie supérieure gauche de la concordance. Un ensemble de boutons permettant la navigation au travers de la pagination des résultats est disponible à la fois dans la partie supérieur droite et inférieure droite. Chaque page contient 100 éléments.

Par défaut, les résultats ne sont pas triés (correspond à l'ordre des textes dans la base de données). Si plusieurs occurrences proviennent d'un même texte, elles seront affichées successivement dans la liste des résultats (l'identifiant du texte « BDLC#XXX » affiché en début de ligne permet de les repérer aisément). Cela ne sera pas nécessairement le cas si des options de tri sont spécifiées.

Après tout changement au niveau des options de tri, il est nécessaire de relancer la requête pour que le résultat soit réactualisé.

L'affichage (ou le masquage) des métadonnées et du contexte étendu relatif à une occurrence en particulier est obtenu en cliquant sur l'occurrence en question (illustration 8).

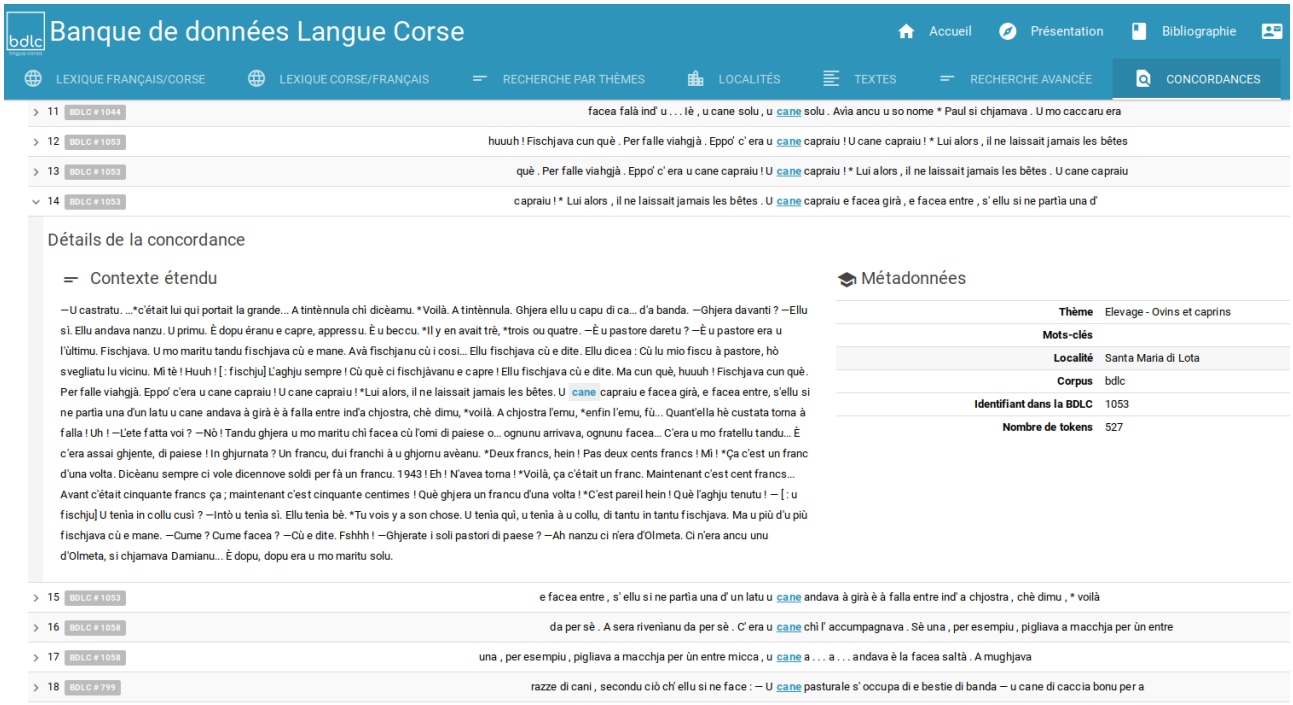

Illustration 8: Concordance avec affichage du contexte étendu et des métadonnées relatives à une occurrence en particulier