

Manuel de TXM



Version 0.8.0

Juillet 2019

Extension Annotation URS (Unité-Relation-Schéma) version 1.0

Copyright © 2019 - projet ANR Democrat

Copyright © 2013-2019 - ENS de Lyon & Université de Franche-Comté

Copyright © 2011-2012 - ENS de Lyon - <http://textometrie.ens-lyon.fr>

Copyright © 2008-2010 - projet ANR Textométrie

Éditeur : Serge Heiden (ENS de Lyon, coordinateur développement de la plateforme TXM)

Contributeurs (par ordre alphabétique) : Matthieu Decorde, Céline Guillot-Barbance, Serge Heiden, Frédéric Landragin, Alexei Lavrentiev, Frédérique Mélanie-Becquet, Bruno Oberlé, Bénédicte Pincemin, Matthieu Quignard, Francesca Rebasti, Lucie Rousier-Vercreuysen, Zeina Tmart.



Ce document est publié sous licence

Creative Commons BY-NC-SA :

<http://creativecommons.org/licenses/by-nc-sa/3.0/fr>

Table des mises à jour (lignes plus anciennes à la fin du document)

| | | |
|------------|----------|---|
| 11/02/2019 | SH,MD,SJ | Màj pour TXM 0.8.0, dont outils d'annotation |
| 03/07/2019 | SH,MD | Màj section 7.2 Annotation avec un modèle Unité-Relation-Schéma (URS) au fil du texte |

N° d'édition : 1461

Contenu : 34 pages, 8541 mots, 2 ill., 2 tab.

Date d'édition : 03/07/19, 18:23:16

7.2 Annotation avec un modèle Unité-Relation-Schéma (URS) au fil du texte

Les outils d'annotation Unité-Relation-Schéma sont accessibles par le biais de l'extension TXM appelée « Annotation URS (Unité-Relation-Schéma) ». Ils fonctionnent avec tous les corpus de TXM.

Dans un modèle d'annotation Unité-Relation-Schéma (URS), les **Unités** (ou entités) portent sur une *séquence contiguë de mots*. Elles possèdent un nombre quelconque de *propriétés* et peuvent être reliées entre elles par les deux autres types d'éléments. Les **Relations** quant à elle *relient* deux éléments du modèle, et ont un nombre quelconque de propriétés (relation de type 1-à-1). Enfin, les **Schémas** *contiennent* des éléments du modèle, et ont un nombre quelconque de propriétés (relation de type 1-à-n).

Les annotations URS ont été utilisées pour encoder des éléments du discours comme les chaînes de co-référence au sein des textes (Schnedecker, Glikman, & Landragin, 2017), qui sont composées d'unités de type « mention » contenues dans des schémas de type « chaîne de co-référence ». Mais l'utilisateur peut définir n'importe quel modèle d'annotation - basé sur des unités, relations et schémas - en définissant une « structure d'annotation » spécifique à un corpus. La structure d'annotation par défaut de tous les corpus TXM est composée d'unités de type « entity » ayant une seule propriété « property ».

Le modèle URS a été défini et implémenté à l'origine dans le logiciel Glozz (Widlöcher & Mathet, 2009). Il est documenté dans le manuel « Glozz User's Manual » <http://glozz.free.fr/glozzManual_1_0.pdf>. Il a également été implémenté dans le logiciel Analec¹ (Landragin, Poibeau, & Victorri, 2012).

L'annotation d'éléments d'un modèle URS dans TXM se fait par le biais de l'extension « Annotation URS »² qui reprend l'implémentation du logiciel Analec. Elle permet d'annoter les textes d'un corpus et de réaliser diverses exploitations de ces annotations.

La première version de l'extension permet d'annoter interactivement les unités au sein des éditions de texte de TXM, d'enrichir l'annotation par commandes ou par macros, de vérifier sa cohérence et de procéder à diverses extractions pour affichage ou décomptes.

L'interface d'annotation des unités reproduit celle du logiciel Analec, dont nous vous invitons à consulter le manuel <http://www.lattice.cnrs.fr/IMG/pdf/ManuelAnalec_1501.pdf> pour apprendre à utiliser la barre d'outils d'annotation des unités (en particulier la section “Annoter des unités” page 15).

7.2.1 Installation de l'extension Annotation URS

Appeler la commande « Fichier > Ajouter une extension » :

1. choisir l'extension « Annotation URS (Unités-Relations-Schemas) » ;
2. accepter la licence ;
3. lancer l'installation ;
4. redémarrer TXM.

¹ <http://lattice.cnrs.fr/Telecharger-Analec>

² Le développement a été financé par le projet ANR DEMOCRAT <<http://www.lattice.cnrs.fr/democrat>>.

7.2.1.1 Compatibilité et Prérequis

En mars 2017 sous Windows 7 : pour que l'extension Annotation URS puisse mettre en évidence les unités annotées dans l'édition vous devez disposer d'une version d'Internet Explorer plus récente que celle livrée avec le système d'exploitation (dans ce cas il faut mettre à jour votre Internet Explorer vers une version plus récente) [voir le ticket [#2017](#) <<http://forge.cbp.ens-lyon.fr/redmine/issues/2017>>].

7.2.2 Modèle d'annotation URS par défaut

Tout corpus importé dans TXM peut être annoté avec un modèle d'annotation URS par défaut élémentaire : ce modèle est constitué d'unités de type « entity » ayant une seule propriété « property ».

Pour annoter avec un modèle d'annotation URS personnalisé, on peut :

- soit modifier directement le modèle d'annotation au moyen de la commande « URS > Éditer la structure d'annotation »³ ;
- soit importer la description d'un modèle d'annotation depuis un fichier d'extension « .aam » à l'aide de la commande « URS > Importer un modèle Glozz... ».

7.2.3 Importer un corpus déjà annoté avec Analec ou Glozz

Il y a deux façons d'importer dans TXM un corpus déjà annoté dans le logiciel Analec :

- Import XML-TEI Analec de corpus : à partir d'un fichier XML-TEI Analec exporté depuis Analec ;
- Import Glozz de corpus : à partir de 3 fichiers d'extensions « .aa », « .aam » et « .ac » exportés depuis Analec ou Glozz.

7.2.3.1 Import de corpus XML-TEI Analec

- lancer la commande « URS > Import XML-TEI Analec Corpus... » en lui fournissant l'argument suivant :
 - xmlFile : le fichier XML-TEI Analec exporté depuis Analec (eg Le_Capitaine_Fracasse_or.xml)
→ un nouveau corpus LECAPITAINEFRACASSEOR est ajouté à la vue Corpus (il a été importé par une version interne du module TXT+CSV de TXM et contient les annotations Analec et leur modèle URS)

Le corpus résultat contient un seul texte dont le nom correspond à celui du fichier importé.

Exemple de fichier XML-TEI Analec du projet ANR DEMOCRAT : [Le_Capitaine_Fracasse_or.xml](#).

7.2.3.2 Import de corpus Glozz : à partir de trois fichiers .aa, .aam et .ac

- lancer la commande « URS > Import a Glozz corpus... » en lui fournissant les arguments suivants :
 - **aafile** : le fichier d'annotations (eg Cleves-brut-simplifie-annotations.aa)

³ le menu principal « URS » ne s'affiche que lorsqu'un corpus est sélectionné dans la vue Corpus.

- **aamfile** : le fichier modèle d'annotations (modèle URS) (eg Cleves-brut-simplifie-structure-annotation.aam)
- **acfile** : le fichier texte (eg Cleves-brut-simplifie-texte.ac) → un nouveau corpus CLEVESBRUTSIMPLIFIETEXTE est ajouté à la vue Corpus (il a été importé par une version interne du module TXT+CSV de TXM et contient les annotations Analec ou Glozz et leur modèle d'annotation URS)

Exemple de fichiers .aa, .aam et .ac exportés depuis Analec pour le projet ANR DEMOCRAT : [Cleves-brut-simplifie.zip](#).

Le module d'import vous demandera de désigner un répertoire contenant votre corpus (.ac) et vos annotations (.aa). Le fichier modèle d'annotation (.aam) peut se trouver dans un autre répertoire.

Le corpus résultat contient un seul texte dont le nom correspond à celui du fichier d'extension « .ac ».

7.2.4 Importer des annotations URS

Il est possible d'importer dans un corpus TXM pré-existant des annotations URS représentées dans le format XML-TEI URS. Pour cela il faut que les textes du corpus contiennent des mots correspondants aux annotations (les annotations sont définies par les empan de mots sur lesquels elles s'appliquent et les mots sont repérés par leur identifiant).

La commande “Import XML-TEI URS annotations...” du menu principal URS importe les annotations URS à partir d'un ou de plusieurs fichiers d'annotations encodées au format XML-TEI URS.

Elle prend comme paramètres :

- **directory** : un répertoire contenant les fichiers d'annotations encodées au format XML-TEI URS (ce qui correspond au répertoire « annotations » d'une archive « .urs ») ;
- **aamFile** : un fichier au format XML Glozz '.aam' de déclaration de la structure d'annotation utilisée par les fichiers d'annotations (ce qui correspond au fichier « .aam » d'une archive « .urs ») ;
- **resetAnnotations** : une option pour supprimer ou non toutes les annotations pré-existantes du corpus avant l'import.

Remarques :

- les textes composant le corpus recevant les annotations doivent contenir des mots dont l'identifiant correspond à celui des mots sur lesquels sont ancrées les annotations importées ;
- les annotations importées ne trouvant pas de mots avec leurs identifiants cibles dans le corpus seront ignorées ;

lors de l'import à partir de N fichiers d'annotations correspondant à N textes présents dans le corpus, les annotations seront importées automatiquement dans leur texte à partir de leurs identifiants de mots.

7.2.5 Annoter des unités interactivement depuis une édition de texte

Soit le corpus LAFAYETTECLEVES produit par le projet Democrat, à télécharger depuis <<https://sharedocs.huma-num.fr/wl/?id=2gYtlk2JTyEPc9ICaeIwXTbbAukmrOHK>>, et chargé dans TXM.

7.2.5.1 Lancer une session d'annotation

- ouvrir l'édition du corpus, par exemple LAFAYETTECLEVES, (clic droit sur l'icone du corpus et menu contextuel « Édition ») ;
- cliquer sur le bouton « Annoter » (bouton crayon situé en bas de l'édition à gauche) ;
- la barre d'outils des unités s'ouvre en haut de l'édition ;
- ainsi que la fenêtre d'édition des unités (vue « Unit ») située en bas de l'édition ;
- remarque : dans TXM les annotations URS sont posées sur les mots et non sur les caractères comme dans Analec. Les mots sont les mots simples définis par TXM (pas composés, souvent étiquetés et lemmatisés, etc.) ou par l'utilisateur selon le module d'import de textes sources qui a été utilisé pour créer le corpus. Le corpus LAFAYETTECLEVES a été importé avec le module TXT+CSV en appliquant le modèle TreeTagger français (le s long - f - ruine les performances). Il s'agit donc de mots standards (par défaut) de TXM. La ponctuation est assimilée aux mots (eg la virgule est un mot que l'on peut sélectionner).

7.2.5.2 Visualiser les unités présentes

- aller à la page contenant les premiers mots du texte, par exemple la page 2 de l'édition de LAFAYETTECLEVES
- dans la barre d'outils des unités sélectionner un type d'unité → toutes les unités de ce type sont mises en évidence dans la page.

7.2.5.3 Créer des unités

- sélectionner quelques caractères ou la totalité d'un mot dans l'édition
 - créer l'unité correspondant au mot avec le bouton « Créer » de la barre d'outils des unités ou la touche « Entrée » du clavier
- sélectionner plusieurs mots (ou portions de mots) dans l'édition
 - créer l'unité avec le bouton « Créer »
- double-cliquer sur un mot
 - créer l'unité avec la touche « Entrée » du clavier
- les unités sont mises en évidence avec la couleur vert clair et l'unité courante avec du vert foncé

7.2.5.4 Éditer les propriétés d'une unité

- cliquer directement sur l'unité ou bien la sélectionner par son identifiant dans le menu des identifiants d'unités (voir la section « sélection des unités » ci-dessous) ;

- les propriétés de l'unité s'affichent dans la vue « Propriétés » dont l'onglet est nommé d'après le nom du corpus, le type d'unité et son adresse en mots dans le corpus ;
- la valeur d'une propriété peut être soit modifiée à partir d'une liste de valeurs déjà connues (bouton [▼]) ;
- soit saisie directement :
 - le raccourcis clavier Ctrl-ESPACE active le mécanisme d'auto-complétion
 - l'auto-complétion affiche la liste des valeurs déjà connues contenant (n'importe où) ce qui a déjà été saisi
 - cette liste sert à choisir directement la valeur souhaitée sans avoir à la re-saisir entièrement
 - navigation dans la liste
 - sélectionner la valeur suivante ou précédente (+1 ou -1)
 - Flèche_vers_le_bas ou Flèche_vers_le_haut
 - sélectionner la valeur +10 ou -10
 - Page_vers_le_bas ou Page_vers_le_haut
 - sélectionner la première ou la dernière valeur
 - Début ou Fin
 - l'auto-complétion peut également afficher la liste des valeurs déjà connues commençant par ce qui a déjà été saisi (mode 'préfixe')
 - le choix du mode d'auto-complétion, recherche par 'infixe' ou par 'préfixe', est réglable dans la préférence « TXM > Utilisateur > Annotation URS > Auto-complétion »
 - continuer la saisie réduit la liste des valeurs proposées
 - la touche Échap ou Esc (en haut à gauche du clavier) permet de quitter le mode d'auto-complétion : la saisie continue là où elle en était

7.2.5.5 Sélectionner des unités

- depuis l'édition :
 - cliquer sur l'unité → l'unité est mise en évidence et ses propriétés s'affichent dans la vue Unit ;
 - aller à l'unité suivante avec le raccourci clavier Ctrl-Flèche_vers_le_bas (tout en maintenant la touche 'Ctrl' enfoncée, appuyer sur la touche 'Flèche_vers_le_bas') ;
 - aller à l'unité précédente avec le raccourci clavier Ctrl-Flèche_vers_le_haut.
- depuis la barre d'outils des unités (située en haut de l'édition)
 - utiliser les boutons de flèches droite [►] et gauche [◀] pour aller à l'unité suivante ou précédente
 - cliquer sur l'identifiant courant :
 - aller à l'unité suivante avec la touche Flèche_vers_le_bas ou avec la molette de la souris vers le bas ou deux doigts glissés vers le haut sur le trackpad (le sens peut être inversé selon les systèmes d'exploitation) ;
 - aller à l'unité précédente avec la touche Flèche_vers_le_haut ou avec la molette de la souris vers le haut ou deux doigts glissés vers le bas sur le trackpad ;
 - aller 10 unités plus loin avec la touche Page_vers_le_bas ;
 - aller 10 unités en arrière avec la touche Page_vers_le_haut ;
 - la navigation dépassant la fin de la liste cycle au début de la liste (même chose pour le dépassement du début de la liste) ;

- sélectionner l'unité par la saisie de son identifiant puis validation avec la touche Entrée :
 - pendant la saisie, on peut activer l'auto-complétion avec le raccourcis Ctrl-Espace ;
 - quand l'auto-complétion est activée la liste des identifiants correspondants à ce qui a déjà été saisi s'affiche
 - cette liste sert à choisir directement l'identifiant souhaité sans avoir à saisir la totalité de l'identifiant
 - navigation dans la liste
 - sélectionner l'identifiant suivant ou précédent (+1 ou -1)
 - molette de la souris
 - Flèche_vers_le_bas ou Flèche_vers_le_haut
 - sélectionner l'identifiant +10 ou -10
 - Page_vers_le_bas ou Page_vers_le_haut
 - continuer la saisie réduit la liste des identifiants proposés

7.2.5.6 Rechercher des unités par la valeur de leurs propriétés

On peut rechercher des unités par leurs propriétés en cliquant sur le bouton “Chercher” (icône de loupe). Cette commande ouvre un formulaire de recherche dans une nouvelle vue, qui s'ouvre par défaut en bas de l'interface de TXM.

Le formulaire de recherche comporte :

- sur la première ligne :
 - à gauche
 - un bouton “Chercher” qui lance la recherche en utilisant les critères courants du formulaire ;
 - un bouton de remise à zéro des critères de recherche.
 - à droite
 - des boutons de navigation dans les résultats : aller au premier résultat, précédent, numéro de résultat courant, suivant, dernier ;
 - un bouton “Concordance” pour afficher la concordance des mots des unités correspondants à la recherche.
- sur les lignes suivantes : un champ de recherche par propriété
 - on saisit la valeur recherchée dans le champ de la propriété concernée ;
 - chaque champ dispose d'un menu déroulant des valeurs possibles de la propriété ;
 - un champ peut contenir soit une valeur exacte soit une expression régulière ;
 - on peut faire la recherche dans plusieurs propriétés à la fois pour une recherche combinée ;
 - si un champ est laissé vide alors la propriété ne participe pas à la recherche.

7.2.5.7 Rectifier les bornes d'une unité

Il y a 3 modes de rectification des bornes d'une unité :

- corriger la borne gauche :
 - 1) sélectionner l'unité

- 2) cliquer sur le bouton 'corriger la borne gauche' ”[↔”
- 3) cliquer sur le mot qui sera la nouvelle borne gauche de l'unité
- corriger la borne droite :
 - 1) sélectionner l'unité
 - 2) cliquer sur le bouton 'corriger la borne droite' “↔]”
 - 3) cliquer sur le mot qui sera la nouvelle borne droite de l'unité
- corriger simultanément la borne gauche et la borne droite :
 - 1) sélectionner l'unité
 - 2) cliquer sur le bouton 'corriger les bornes gauche et droite' ”[↔”
 - 3) sélectionner les mots qui formeront les nouvelles bornes de l'unité (comme lors de la création d'une unité)

7.2.5.8 Créer des unités à cheval sur deux pages d'édition

Selon le module d'import utilisé, et certains paramètres d'import comme le nombre de mots par page, certaines éditions de textes peuvent être paginées. Dans ce cas s'il est nécessaire de créer des unités à cheval sur plusieurs pages :

- créer l'unité à partir de ses premiers mots dans la première page (derniers mots de la page) ;
- lancer la rectification des bornes ;
- passer à la page suivante ;
- étendre l'unité à ses derniers mots
→ actuellement l'affichage bugue : les unités s'effacent, on peut changer de page et revenir pour ré-afficher correctement ;
- on procède de la même manière pour créer une unité à partir de ses derniers mots situés dans la deuxième page (premiers mots de la page) ;
- si l'unité s'étend sur plus de deux pages, suivre la même procédure en se déplaçant du nombre de pages nécessaire.

7.2.5.9 Supprimer une annotation

- sélectionner une unité en vert clair par un clic, elle devient vert foncé ;
- la supprimer avec le bouton “Supprimer” ou la touche “Suppr” du clavier.

7.2.6 Enregistrer les annotations

Le travail d'annotation est sauvegardé avec le bouton d'enregistrement (bouton crayon+disquette situé en bas à gauche de l'édition) ou bien avec le raccourcis clavier Ctrl-S.

Il est important de sauvegarder régulièrement son travail, notamment parce que la fonction annuler n'est pas disponible.

7.2.7 Annoter, Vérifier, Exploiter, Exporter par commandes

Ces commandes sont accessibles depuis le menu principal ou contextuel “URS > Outils” ou bien depuis le répertoire “URS” de la vue Macro en ce qui concerne les macros Democrat.

7.2.7.1 Paramètres généraux

Toutes les commandes URS utilisent des paramètres communs, renseignés en début de formulaire de paramètres.

La documentation de chaque commande précise, le cas échéant, les paramètres qui lui sont spécifiques.

7.2.7.1.1 Corpus, Sous-corpus & Partition

Toutes les commandes URS s'appliquent à une sélection de corpus, de sous-corpus ou de partition.

Si la sélection est multiple, les commandes produisent un résultat contrastif entre les objets sélectionnés (en tableau, par exemple UnitsIndex) ou bien itèrent sur les objets sélectionnés (par exemple UnitsList).

Dans le cas des partitions, les commandes travaillent sur les sous-corpus correspondant à leurs parties.

7.2.7.1.2 Sélection d'éléments URS

Les commandes utilisent des paramètres généraux de sélection d'éléments URS (unités et schémas dans la version 1.0) :

- les éléments peuvent être sélectionnés par :
 - a) requête d'unités (section B)
 - b) requête de schéma (section A)
 - c) une combinaison de requête d'unités et de schéma (section A & B)
 - dans ce cas, la sélection d'unités se limite aux unités des schémas sélectionnés (si la sélection de schéma ne retourne aucun schéma, aucune unité ne peut être sélectionnée), ou bien la sélection de schémas se limite à ceux contenant les unités sélectionnées
 - quand la sélection de schémas n'est pas renseignée, la sélection d'unités se fait sur la totalité des unités du corpus
- A) sélection de **schémas** :
 - par propriétés :
 - la sélection par champs (b) est alternative à la sélection par requête URS (a) (voir la section **Requêtes URSQL** qui suit)
 - a) par une requête URSQL de schéma de la forme TYPE@PROPRIETE=VALEUR
 - *schema_ursql* : requête de sélection de schéma (exemple 'CHAINE@REF=Jean.*')
 - b) par une combinaison de champs (si un des champs n'est pas renseigné alors il est ignoré) :
 - *schema_type* : sélection par type (exemple 'CHAINE')
 - *schema_property_name* : sélection par propriété (exemple 'REF')

- *schema_property_value* : expression régulière de sélection de valeurs de propriété (par défaut '.'*' pour sélectionner toutes les valeurs. Exemple 'Jean.*')
- par longueur :
 - *minimum_schema_size* : longueur minimale en unités du corpus sélectionné des schémas à prendre en compte dans le calcul (utile notamment pour filtrer les singletons)
 - *maximum_schema_size* : longueur maximale en unités du corpus sélectionné des schémas à prendre en compte dans le calcul
- B) sélection d'**unités** :
 - par propriétés :
 - par une requête URSQL d'unité de la forme TYPE@PROPRIETE=VALEUR
 - *unit_ursql* : requête de sélection d'unité (exemple 'MENTION@CATEGORIE=GN.*')
 - par une combinaison de champs (si un des champs n'est pas renseigné alors il est ignoré) :
 - *unit_type* : sélection par type (exemple 'MENTION')
 - *unit_property_name* : sélection par propriété (exemple 'CATEGORIE')
 - *unit_property_value* : expression régulière de sélection de la valeur de la propriété (par défaut '.'*' pour sélectionner toutes les valeurs) (exemple 'GN.*')
 - par position dans les schémas ou dans les structures CQP
 - *position_in_schema* : position de l'unité, dans l'ordre du texte dans le schéma
 - exemples :
 - 1 : première unité du schéma
 - 2 : deuxième unité du schéma
 - etc.
 - 0 : toutes les unités du schéma (pas de sélection par position dans le schéma)
 - -1 : dernière unité du schéma
 - -2 : avant-dernière unité du schéma
 - etc.
 - *cql_limit* : expression CQL désignant les limites de structures CQP à utiliser dans le calcul. L'expression doit mobiliser une structure disponible dans le corpus et faire un 'expand to' pour obtenir autant d'intervalles que de structures souhaitées

- par exemple :
 - `<div> []` expand to div, pour utiliser toutes les limites de div
 - `<div_type="chapitre"> []` expand to div, pour utiliser toutes les limites de div de type “chapitre”
 - `<p> []` expand to p, pour utiliser toutes les limites de paragraphes
- remarques :
 - la macro `cqp/ListMatches` peut être utile pour vérifier les positions des limites, les matches, d'un sous-corpus. Par exemple un sous-corpus créé en mode avancé avec la requête `cql_limit`. Le sous-corpus correspondant est constitué de la liste de toutes les séquences de mots correspondants à toutes les occurrences de la requête dans le corpus. Chaque occurrence est délimitée par la position de son premier mot (début) et par la position de son dernier mot (fin) dans le corpus. La position du premier mot d'un corpus est par convention 0 (et non 1).
 - la macro `commands/StructureIndex` peut être utile pour vérifier les positions des limites de structures
- *position_in_matches* : position de l'unité, dans l'ordre du texte dans les matchs de la requête `cql_limit`
 - exemples :
 - 1 : première unité du match
 - 2 : deuxième unité du match
 - etc.
 - 0 : toutes les unités du match (pas de sélection par position dans la structure)
 - -1 : dernière unité du match
 - -2 : avant-dernière unité du match
 - etc.
- *strict_inclusion* : tous les mots de l'unité doivent se situer au sein du sous-corpus ou de la structure définie par `cql_limit` pour que l'unité puisse être sélectionnée

7.2.7.1.2.1 Requêtes URSQL

La sélection sur critères d'éléments URS peut s'exprimer par le biais d'un certain nombre de champs dans un formulaire de paramètres ou bien par le biais d'une expression unique utilisant le langage de requête URSQL. Le mécanisme de sélection par requêtes URS, plus concis, remplace petit à petit la sélection par champs.

Une requête permet de sélectionner **des** éléments par leur type, par une propriété ou par une valeur de propriété.

C'est le contexte d'utilisation d'une requête URSQL qui détermine le type d'éléments concerné par la sélection : Unité, Schéma ou Relation. Par exemple on trouve des champs de requêtes URSQL pour sélectionner des Unités ou pour sélectionner des Schémas.

Une requête est de la forme « TYPE@PROPRIÉTÉ=VALEUR_DE_PROPRIÉTÉ ».

Chacun des 3 critères, TYPE, PROPRIÉTÉ et VALEUR_DE_PROPRIÉTÉ (identifiants en majuscules), peut être exprimé sous la forme d'une expression régulière.

Si on souhaite ignorer un critère de sélection, on ne le renseigne pas.

Syntaxes possibles :

- « » (champ vide ou chaîne vide) : aucun critère de sélection n'est appliqué, tous les éléments sont sélectionnés
- TYPE : on sélectionne tous les éléments d'un certain TYPE
- TYPE@PROPRIÉTÉ : on sélectionne tous les éléments d'un certain TYPE et ayant une certaine PROPRIÉTÉ
- TYPE@PROPRIÉTÉ=VALEUR_DE_PROPRIÉTÉ : on sélectionne tous les éléments d'un certain TYPE et ayant une certaine PROPRIÉTÉ à une certaine VALEUR_DE_PROPRIÉTÉ
 - exemple, dans le corpus PERRAULTCONTES2 :
Mention@REF=Anne
- @PROPRIÉTÉ=VALEUR_DE_PROPRIÉTÉ : on sélectionne tous les éléments ayant une certaine PROPRIÉTÉ à une certaine VALEUR_DE_PROPRIÉTÉ (pour n'importe quel TYPE)
- @PROPRIÉTÉ : on sélectionne tous les éléments ayant une certaine PROPRIÉTÉ (pour n'importe quel TYPE)
- =VALEUR_DE_PROPRIÉTÉ : on sélectionne tous les éléments ayant une certaine VALEUR_DE_PROPRIÉTÉ (pour n'importe quel TYPE et n'importe quelle PROPRIÉTÉ)

7.2.7.1.3 Propriétés d'éléments à utiliser

Les commandes utilisent des paramètres généraux de désignation de propriétés d'éléments URS :

- *unit_property_display* : nom de la propriété d'unité à utiliser et à afficher
 - si la propriété n'existe pas, certains calculs peuvent utiliser la propriété CQP correspondante des mots recouverts par l'unité (si elle existe). Par exemple : 'flemma'
- *schema_property_display* : nom de la propriété de schéma à utiliser et à afficher

7.2.7.1.4 Autres

- *debug* : active l'affichage de messages de mise au point

7.2.7.2 Commandes d'annotation

Ces commandes, accessibles depuis le menu principal ou contextuel “URS > Outils > Annotation”, sont complémentaires des outils d'annotation interactifs accessibles depuis l'interface utilisateur des éditions de textes. Elles permettent, par exemple, de :

- créer des unités par lots à partir de la sélection de séquences de mots par requêtes CQL
- projeter rapidement des annotations sur des lots d'unités sélectionnées par requêtes URSQL
- supprimer des unités par lots

7.2.7.2.1 Concordance To Units

Créer des unités à partir des pivots d'une concordance :

- la concordance est sélectionnée par son icône dans la vue Corpus ou par sa fenêtre de résultats activée (ayant le focus de la souris)
- *unit_type* : type des unités à créer
- *move_start* : décale la position du début de l'unité de la valeur indiquée. Exemple : “1” avance la position d'un seul mot ; “-1”, recule la position d'un mot (0 par défaut)
- *move_end* : décale la position de la fin de l'unité de la valeur indiquée. Exemple : “1” avance la position d'un seul mot ; “-1”, recule la position d'un mot (0 par défaut)
- *create_only_if_new* : crée une unité seulement si le pivot n'a pas déjà une annotation de type *unit_type*
- *property_name* : propriété des unités à ajouter (optionnel)
- *property_value* : valeur de la propriété à ajouter (optionnel)

7.2.7.2.2 Units Annotate

Affecter ou modifier une propriété des unités sélectionnées.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*

¹⁾Voir la section Sélection d'éléments URS pour leur description

- sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql*
 - par position
 - *position_in_schema*
 - *cql_limit*
 - *position_in_matches*
 - *strict_inclusion*

Paramètres spécifiques :

- *property_name* : propriété à modifier
- *property_value* : valeur à affecter

7.2.7.2.3 Units Delete

Supprimer des unités sélectionnées.

Remarque : **attention** si aucun critère de sélection n'est fourni (en schémas et en unités), toutes les unités sont supprimées.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql*
 - par position

- *position_in_schema*
- *cql_limit*
- *position_in_matches*
- *strict_inclusion*

Paramètres spécifiques :

- pas de paramètres spécifiques

7.2.7.2.4 Reset All Annotations

Supprimer **toutes** les annotations : unités, relations et schémas, du corpus :

- *CONFIRMATION* : cocher pour confirmer la suppression

7.2.7.3 Commandes de vérification de la cohérence des annotations

Il est fortement conseillé d'appliquer systématiquement ces commandes, accessibles depuis le menu principal ou contextuel “URS > Outils > Vérification”, avant d'utiliser des commandes d'exploitation. En effet, la qualité des calculs de ces dernières est très sensible aux manques et aux redondances d'annotations d'un corpus.

7.2.7.3.1 Check Annotation Structure Values

Vérifier que toutes les valeurs d'une propriété définie dans la structure d'annotation sont effectivement utilisées par les annotations :

- *unit_type* : type d'unité à vérifier
- *unit_property* : propriété à vérifier
- *pruneUnusedValues* : supprimer les valeurs non-utilisées de la structure d'annotation

7.2.7.3.2 Check Duplicates In Schema

Lister pour chaque schéma les unités ayant les mêmes propriétés aux mêmes valeurs.

7.2.7.4 Commandes d'exploitation d'unités (lister, compter, visualiser...)

Commandes accessibles depuis le menu principal ou contextuel “URS > Outils > Exploitation”.

Des unités doivent avoir été annotées dans le corpus pour que ces commandes fonctionnent. Les sélections par schémas nécessitent également que des schémas aient été annotés dans le corpus.

7.2.7.4.1 Units Summary

Dénombrer les unités sélectionnées ainsi que les valeurs de leurs propriétés.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql*
 - par position
 - *position_in_schema*
 - *cql_limit*
 - *position_in_matches*
 - *strict_inclusion*

Paramètres spécifiques :

- pas de paramètres spécifiques

7.2.7.4.2 Units Index

Index hiérarchique d'une propriété des unités sélectionnées.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*

- sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql*
 - par position
 - *position_in_schema*
 - *cql_limit*
 - *position_in_matches*
 - *strict_inclusion*

Paramètres spécifiques :

- *unit_property_display* : propriété d'unité à utiliser pour le calcul, par exemple 'CATEGORIE'
- *word_property_display* : propriété de mot à utiliser pour le calcul, par exemple 'frpos' (si ce paramètre est renseigné, le paramètre *unit_property_display* est ignoré)
- *output_showlegend* : afficher la légende du diagramme à barres
- *output_fmin* : seuils en fréquence des valeurs à afficher
- *output_histogram* : afficher le diagramme à barres
- *output_lexicaltable* : créer une table lexicale à partir des données de l'index

7.2.7.4.3 Units List

Afficher la liste des unités sélectionnées avec différents modes d'affichage.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- sélection d'unités
 - par propriétés

- par requête
 - *unit_ursql*
- par position
 - *position_in_schema*
 - *cql_limit*
 - *position_in_matches*
- *strict_inclusion*

Paramètres spécifiques :

- *output_mode* : mode d'affichage des résultats
 - FORMATED (par défaut) : affiche les informations des unités sous la forme : #n°, adresse du premier mot - adresse du dernier mot, formes de surface, propriétés
 - TABULATED : affiche les informations des unités séparées par des tabulations : #n°, adresse premier mot - adresse dernier mot, formes de surface, propriétés
 - COUNT : affiche seulement le décompte d'unités
 - CQL : affiche une requête CQL sélectionnant toutes les séquences de mots des unités
 - Concordance : crée la concordance des séquences de mots des unités

7.2.7.4.4 Units Interdistance

Afficher les moyennes, médianes, quartiles et histogrammes des distances et des cadences entre les unités sélectionnées.

La distance est calculée en mots : depuis le dernier mot d'une unité jusqu'au premier mot de l'unité suivante dans le corpus.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- sélection d'unités

- par propriétés
 - par requête
 - *unit_ursql*

Paramètres spécifiques :

- pas de paramètres spécifiques

7.2.7.4.5 Units Referential Density

Calculer le rapport entre le nombre d'unités d'un certain type et le nombre de mots du corpus (en %).

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql*
 - par position
 - *position_in_schema*
 - *cql_limit*
 - *position_in_matches*
 - *strict_inclusion*

Paramètres spécifiques :

- pas de paramètres spécifiques

7.2.7.4.6 Units Referential Stability

Calculer le rapport entre le nombre d'unités sélectionnées et le nombre de formes différentes utilisées par ces unités (taille du vocabulaire des unités) dans le corpus.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql*

Paramètres spécifiques :

- *word_property* : propriété de mots CQP à utiliser, Exemple 'word' pour la forme graphique des mots
- *show_values* : afficher les valeurs de la propriété de mots *word_property* (peut être assez long)

7.2.7.5 Commandes d'exploitation de schémas (lister, compter, visualiser...)

Commandes accessibles depuis le menu principal ou contextuel “URS > Outils > Exploitation”.

Des schémas doivent avoir été annotés dans le corpus pour que ces commandes fonctionnent. Les sélections par unités nécessitent également que des unités aient été annotées dans le corpus.

7.2.7.5.1 Schemas Summary

Afficher le nombre de schémas sélectionnés, la longueur moyenne et médiane (en unités) des schémas sélectionnés, ainsi que le nombre de valeurs différentes de leurs propriétés.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur

- *minimum_schema_size*
- *maximum_schema_size*

Paramètres spécifiques :

- pas de paramètres spécifiques

7.2.7.5.2 Schemas Coverage

Ajoute la propriété « LOCALISATION » aux schémas sélectionnés en y enregistrant la référence de la plus petite structure CQP contenant l'empan du schéma.

La référence est de la forme « structure_propriété valeur » (par exemple “p_n 1”). Elle contient « N/A » si le schéma ne dépasse pas le seuil *minimum_schema_size*.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- par sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql*
 - *strict_inclusion*

Paramètres spécifiques :

- *structure_properties* : liste des propriétés de structure à interroger. Les propriétés sont ordonnées par ordre de taille de structure. Par exemple, « div_n, p_n ». Si aucune des structures n'englobe le schéma, la propriété LOCALISATION est affectée à « text_id » (le texte contenant le schéma).

7.2.7.5.3 Schemas Index

Index hiérarchique d'une propriété des schémas sélectionnés.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas

- par propriétés
 - par requête
 - *schema_ursql*
- par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- par sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql*
 - *strict_inclusion*

Paramètres spécifiques :

- *schema_property_display* : propriété de schéma à utiliser pour le calcul, par exemple 'CATEGORIE'
- *output_fmin* : seuils en fréquence des valeurs à afficher

7.2.7.5.4 Schemas List

Afficher les schémas sélectionnés avec leur nombre, leurs propriétés et la liste de leurs unités.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*

Paramètres de sélection d'éléments à afficher :

- sélection d'unités
 - par propriétés

- par requête
 - *unit_ursql*

Paramètres spécifiques :

- *word_property* : propriété de mots CQP à afficher
- *separator* : séparateur entre les valeurs de la propriété
- *buildCQL* : afficher la requête CQL de recherche des mots de chaque schéma

7.2.7.5.5 Schema Lengths

Afficher la longueur moyenne (en unités) des schémas sélectionnés et l'index hiérarchique des longueurs.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql*

Paramètres spécifiques :

- *schema_property_display* : propriété de schéma à afficher
- *sort_column* : colonne à utiliser pour le tri : “freq” (défaut) ou “len” (ordre inverse)
- *max_lines* : nombre de lignes à afficher dans le tableau

7.2.7.5.6 Schemas Progression

Afficher le graphique de progression des schémas en mettant en valeur les unités sélectionnées (la sélection d'unités ne participe pas à la sélection des schémas).

La légende est triée selon l'ordre d'apparition des schémas au fil du corpus (ordre croissant des positions de la première unité du schéma).

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql*
 - par position
 - *position_in_schema*
 - *cql_limit*
 - *position_in_matches*
 - *strict_inclusion*

Paramètres spécifiques :

- *schema_property_display* : propriété de schéma à afficher dans la légende
- *struct_name* : nom de la structure à afficher sur le graphique de progression
- *struct_prop* : nom de la propriété de structure à afficher sur le graphique de progression
- *line_width* : épaisseur des lignes de la progression
- *cummulative* : affiche la courbe de type cummulative, sinon la densité
- *bande_width* : largeur de bande densité de la progression (densité)
- *show_lines* : affiche les lignes de la courbes (cummulative)
- *show_title* : affiche ou pas le titre de la progression
- *show_legend* : affiche ou pas la légende de la progression

7.2.7.5.7 Schema Ratio

Affiche pour chaque schéma sélectionné, le ratio $\frac{\text{unités sélectionnées}}{\text{unités de marge}}$.

Paramètres de sélection d'éléments URS¹⁾ :

- sélection de schémas
 - par propriétés
 - par requête
 - *schema_ursql*
 - par longueur
 - *minimum_schema_size*
 - *maximum_schema_size*
- sélection d'unités
 - par propriétés
 - par requête
 - *unit_ursql* : requête sélectionnant les unités définissant le numérateur du ratio
- sélection d'unités de marge
 - par propriétés
 - par requête
 - *units_margin_ursql* : requête sélectionnant les unités définissant le dénominateur du ratio

Paramètres spécifiques :

- *schema_property_display* : propriété de schéma à afficher
- *sort_column* : colonne à utiliser pour le tri : “score” (défaut, ordre inverse) ou “name” (la propriété de schéma à afficher)
- *max_lines* : nombre de lignes à afficher dans le tableau

7.2.7.6 Commandes d'export

Commandes accessibles depuis le menu principal ou contextuel “URS > Outils > Export”.

7.2.7.6.1 Export Corpus As

Exporter un corpus TXM au format binaire “.txm” sous un nom différent. Le corpus peut alors être rechargé à côté du premier.

Paramètre :

- *new_name* : nouveau nom du corpus, exemple (MONCORPUS)

7.2.8 Macros 'Democrat'

Ces macros sont accessibles depuis la vue Macro, dans le répertoire “URS” et ses sous-répertoires. Elles ont été développées dans le cadre de la procédure d'annotation du partenaire Lyon pour le projet ANR Democrat et sa structure d'annotation (chaînes de co-références de mentions). En tant que macros, leur algorithme est adaptable directement par l'utilisateur.

La majorité de ces macros a été développée et debuggée par Matthieu Quignard et Alexei Lavrentiev, pour le français moderne et pour l'ancien français.

On utilise ces macros dans les différentes phases du workflow d'annotation :

- Phase 1 (création des mentions)
 - annotation interactive : délimitation et attribution de référent -> MENTION@REF
 - annotation par script : corrige les limites de mentions dans le complément du nom avec la macro **RetoucheComplementDuNom**
 - annotation par script : création de la propriété CATEGORIE avec **Frpos2Cattex** ou **Frpos2Cattex** (médiéval) -> CATEGORIE (dont GN.CHECK, ERREUR, PRO.CHECK)
 - vérification interactive : vérifier que la CATEGORIE est correcte et préciser les valeurs “.CHECK”
 - vérification interactive : finaliser les valeurs “.CHECK”
 - annotation par script : vérification de différents types de cohérences avec **Verifications**
 - vérification interactive : vérifier les MENTION@CHECK!=DONE
- Phase 2 (liste des identifiants de référents fixe)
 - annotation par script : numéroter les référents MENTION@REF=“SI” avec **Si2Singleton**
 - annotation par script : créer des schémas de type CHAINE avec la macro **CreationChaines** -> CHAINE(@REF, @GENRE, @NOMBRE, @NB MAILLONS, @TYPE REFERENT)
 - annotation par script : ajouter la propriété DEFINITUDE aux MENTION avec la macro **AjoutDefinitude** -> MENTION@DEFINITUDE
- Phase 3 (chaque unité est dans la bonne CHAINE)
 - annotation interactive d'unités (vérification)
 - annotation interactive de schémas (vérification et augmentation) [à faire] (slide 34)
- Phase 4 (toutes les CHAINE sont annotées)
 - annotation par script (augmenter) unités [à faire] (slide 37)
 - annotation par script (augmenter) schémas [à faire] (slide 37)

- Phase 5 (unités et schémas fixes)
 - annotation interactive (créer+augmenter) relations [à faire] (slide 39)
- Phase 6 (relations d'anaphore fixes)
 - annotation par script (créer+augmenter) relations (macros **CreationAnaphores**, [à valider])

Documentation des macros par catégories :

7.2.8.1 Vérifier la cohérence des annotations

7.2.8.1.1 Verifications

Cette macro se lance après l'annotation des références et des catégories des mentions. On a besoin de catégories pour repérer les erreurs de type 1, 3, 4 et 7.

Vérifie les erreurs les plus fréquentes :

- 1 repère les mentions sans catégorie : CHECK > CAT
- 2 repère les mentions sans référent : CHECK > REF
- 3 supprime les ponctuations en début et en fin de mention : CHECK > BORNES
- 4 supprime les prépositions autres que 'de' en début de mention : CHECK > BORNES
- 5 supprime automatiquement toutes les mentions vides = sans aucun mot = de longueur 0
- 6 détecte les mentions qui ont exactement les mêmes bornes : CHECK > DOUBLON
- 7 (option) détecte les pronoms hors mention : CHECK > NEW

Paramètres :

- *unit_type* : type des unités à vérifier (MENTION par défaut)
- *pos_property_name* : nom de la propriété catégorie à vérifier (CATEGORIE)
- *cat_name* : nom de la propriété référent à vérifier (REF)
- *ref_name* : nom du champ qui porte l'étiquette morphosyntaxique (type, pos, fpos, fpropos...)
- *checkPronouns* : cocher "checkPronouns" si on veut rajouter les éventuels pronoms oubliés (= hors mention)

Le script ajoute une propriété CHECK en cas d'erreur. Cette propriété indique quel type d'erreur a été rencontré. NB : il peut y avoir plusieurs erreurs pour une même mention.

7.2.8.2 Corriger ou mettre à jour les annotations

7.2.8.2.1 PROPERanaphoriques

Recatégoriser la valeur 'PRO.PER' (pronom personnel) de la propriété CATEGORIE en distinguant les anaphoriques (PRO.PERA) et les déictiques (PRO.PERD).

Paramètres :

- *unit_type* : type d'unité à corriger (exemple 'MENTION')
- *pos_property_name* : propriété de mot à utiliser (exemple 'frlemma')

7.2.8.2.2 RetoucheComplementDuNom

Macro pour corriger une erreur d'annotation : retirer le “de” ou le “d” (forme élidée) du complément du nom.

Paramètres :

- *unit_type* : type d'unité à corriger (exemple 'MENTION')
- *category_name* : propriété d'unité à utiliser (exemple 'CATEGORIE')

Algorithme :

```
POUR CHAQUE MENTION dont le premier mot est "de" (en minuscules)
  SI il existe une autre MENTION dans laquelle celle-ci est totalement incluse
    ALORS incrémenter d'un mot la frontière gauche de la mention
  Ajouter la catégorie CDN.CHECK pour qu'on puisse vérifier facilement le job.
```

Cette macro ne s'appuie pas sur des catégories morphosyntaxiques. On peut donc la faire tourner avant les macros frpos2cattex ou fropos2cattex.

7.2.8.2.3 Si2Singleton

Différencie toutes les valeurs « SI » de la propriété *<ref_property_name>* des unités d'un certain type en les numérotant de 1 à n : « SI_1 », « SI_2 » ... « SI_n ».

Cela permet à la commande CreationChaine de créer n chaînes singletons au lieu de créer une unique chaîne composée de toutes ces unités.

Paramètres :

- *unit_type* : type d'unité à mettre à jour, exemple 'MENTION'
- *ref_property_name* : propriété d'unité à mettre à jour, exemple 'REF'

7.2.8.2.4 SUJ-PARTINF

Recatégorise les mentions étiquetées en SUJ.ZERO pour distinguer les cas des verbes conjugués (qui restent en SUJ.ZERO) des verbes à l'infinitif ou au participe passé/présent qui, eux, vont passer sous la catégorie SUJ.PARTINF.

Cette macro est réversible. Il suffit d'invoquer la macro SearchRemplaceInFile ou SearchRemplaceInDir et substituer CATEGORIE=SUJ.PARTINF en CATEGORIE=SUJ.ZERO.

NB : fonctionne avec le tagset CATTEX pour l'ancien français et avec le tagset TreeTagger pour le français contemporain :

| | CATTEX | TreeTagger |
|------------------|--------|------------|
| Infinitif | VERinf | VER:infi |

| | | |
|--------------------------|--------|----------|
| Participe Passé | VERppa | VER:pper |
| Participe Présent | VERppe | VER:ppre |

Paramètres :

- *unit_type* : type d'unité à corriger (exemple 'MENTION')
- *property_name* : propriété d'unité à corriger (exemple 'CATEGORIE')
- *subj_zero_cat* : valeur de la CATEGORIE pour les sujets zéros (exemple 'SUJ.ZERO')
- *pos_property_name* : propriété de mot à utiliser pour la morphosyntaxe (exemple 'frpos')

7.2.8.2.5 TrimPropertyValues

Élague les caractères espaces et sauts de ligne se trouvant dans les valeurs de propriétés d'éléments URS. Voir les paramètres pour choisir le type d'élagage à apporter.

Cette macro peut être utile après un import d'annotations à partir d'un fichier XML-TEI URS manipulé indépendamment de TXM.

Paramètres :

- *element* : type d'éléments URS à corriger : Unité, Relation ou Schema
- *ursql* : éventuellement sélectionner les éléments à corriger
- *update_value* : remplacer effectivement les valeurs, sinon seules les différences (valeur non élaguée / valeur élaguée) sont affichées
- *trim_spaces* : supprimer les espaces situés en début et en fin de valeur
- *prune_spaces* : remplacer les répétitions d'espaces par un seul
- *prune_newlines* : supprimer les sauts de ligne

7.2.8.3 Annoter

7.2.8.3.1 Unités

7.2.8.3.1.1 Frpos2Categorie

Calculer la valeur de la propriété CATEGORIE des unités à partir des propriétés 'word' (forme graphique) et <pos_property_name> (morpho-syntaxe) des mots qu'elle contient. Le calcul est réalisé pour le français moderne.

Paramètres :

- *unit_type* : type d'unité à annoter, exemple (MENTION)
- *pos_property_name* : propriété de mot pour la morpho-syntaxe, exemple (frpos)
- *reset* : mettre à jour la propriété CATEGORIE même si elle a déjà une valeur

7.2.8.3.1.2 Fropos2Categorie

Calculer la valeur de la propriété CATEGORIE des unités à partir de la propriété <pos_property_name> (morpho-syntaxe) des mots qu'elle contient. Le calcul est réalisé pour l'ancien français.

Paramètres :

- *unit_type* : type d'unité à annoter, exemple (MENTION)
- *pos_property_name* : propriété de mot pour la morpho-syntaxe, exemple (pos)
- *reset* : mettre à jour la propriété CATEGORIE même si elle a déjà une valeur

7.2.8.3.1.3 AjoutDefinitude

Calculer la valeur de la propriété DEFINITUDE, de valeurs 'DEFINI', 'INDEFINI', 'DEMONSTRATIF', 'AMBIGU' ou 'NONE', des unités à partir des propriétés 'word' (forme graphique) des mots qu'elle contient. Le calcul est réalisé pour le français moderne.

Paramètres :

- *unit_type* : type d'unité à annoter, exemple (MENTION)
- *reset* : mettre à jour la propriété DEFINITUDE même si elle a déjà une valeur

7.2.8.3.1.4 Accessibilite

Calculer la valeur de la propriété Accessibilité (accessibilité du référent) des unités à partir de la valeur de leur propriété CATEGORIE.

Paramètres :

- *unit_type* : type d'unité à annoter, exemple (MENTION)

7.2.8.3.1.5 LongueurDesMentions

Calculer la valeur de la propriété LONGUEUR des unités en calculant leur longueur en mots.

Paramètres :

- *unit_type* : type d'unité à annoter, exemple (MENTION)
- *pos_property_name* : propriété de mot à utiliser, exemple (word)
- *reset* : mettre à jour la propriété LONGUEUR même si elle a déjà une valeur

7.2.8.3.1.6 PremierMaillon

Ajouter la propriété NEW à la valeur 'YES' aux unités d'un certain type si la valeur de leur propriété REF est rencontrée pour la première fois et 'NO' sinon.²⁾

Paramètres :

- *unit_type* : type d'unité à annoter, exemple (MENTION)

²⁾ATTENTION : Le champ « NEW » est ajouté dans la structure, mais il n'est pas affiché dans la vue. On ne peut donc pas le corriger... Pour ce faire, il faut d'abord enregistrer les annotations puis relancer TXM. La vue sera alors rafraîchie.

- *reset* : mettre à jour la propriété NEW même si elle a déjà une valeur

7.2.8.3.2 Schémas

7.2.8.3.2.1 CreationChaines

Créer des schémas de type 'CHAINE' contenant les unités ayant une même valeur de propriété REF.

Paramètres :

- *unit_type* : type d'unités à regrouper, exemple (MENTION)
- *ref_property* : propriété d'unité à utiliser pour le référent, exemple (REF)

7.2.8.3.3 Relations

7.2.8.3.3.1 CreationAnaphores

Créer des relations de type 'ANAPHORE' à partir des unités 'MENTION' de schémas 'CHAINE'.

Paramètres :

- *unit_type* : type d'unités à utiliser, exemple (MENTION)
- *schema_type* : type de schémas à utiliser, exemple (CHAINE)

7.2.8.4 Créer des configurations de corpus

7.2.8.4.1 SubCorpus

Crée un sous-corpus du corpus sélectionné dans la vue Corpus à partir d'une requête CQL. Équivalent à la création de sous-corpus en mode avancé.

Paramètres :

- *name* : nom du sous-corpus à créer
- *query* : requête CQL de création du sous-corpus

7.2.8.4.2 BuildSections

Exemple de macro appelant directement la macro SubCorpus plusieurs fois pour créer des sous-corpus correspondant à des chapitres et à certains paragraphes.

Paramètres : aucun. La macro utilise directement le corpus nommé "DIDEROTESSAIS" (le nom du corpus est donc à adapter pour que la macro fonctionne sur un autre corpus).

7.2.8.5 Exporter

7.2.8.5.1 BuildDemocratCorpusVersion1

Préparer le corpus Democrat pour livraison, en mettant à jour les entêtes TEI avec des statistiques et des informations issues d'un tableau de métadonnées fourni en paramètre.

Paramètres :

- *inputDirectory* : répertoire source contenant les fichiers *.txm, teiHeader.xml, metadata.tsv et metadata.properties de départ
 - *.txm : les corpus annotés URS (un texte par corpus)
 - teiHeader.xml : le patron générique d'entête TEI
 - metadata.tsv : les métadonnées de textes
 - metadata.properties : la table de correspondance entre les métadonnées et leur position dans l'entête TEI
- *outputDirectory* : répertoire où sera déposé le résultat
- *debug* : afficher les informations de mise au point

7.2.9 Exporter des annotations

7.2.9.1 Dans un corpus binaire TXM

- S'assurer que les dernières modifications ont bien été enregistrées avec la commande « URS > Sauvegarder les annotations » ;
- Exporter le corpus dans un fichier .txm avec la commande « Fichier > Exporter ».

7.2.9.2 Au format XML-TEI URS

- la commande “Fichier > Exporter > Corpus au format XML-TEI URS...” exporte les annotations dans une archive d'extension « .urs » dont le format est spécifié dans la page de wiki [Commande Exporter Corpus au format XML-TEI URS...](#)⁴ (Grobol, Landragin, & Heiden, 2017).

7.2.9.3 Au format Glozz

- la commande “Fichier > Exporter > Corpus au format Glozz...” exporte les annotations des unités d'un certain type dans un fichier au format Glozz.

⁴https://groupes.renater.fr/wiki/txm-info/public/annotation/specs_annotation_analec/export_import#commande_exporter_corpus_au_format_xml-tei_urs

8 Table des matières

| | |
|---|----|
| 7.2 Annotation avec un modèle Unité-Relation-Schéma (URS) au fil du texte..... | 2 |
| 7.2.1 Installation de l'extension Annotation URS..... | 2 |
| 7.2.1.1 Compatibilité et Prérequis..... | 3 |
| 7.2.2 Modèle d'annotation URS par défaut..... | 3 |
| 7.2.3 Importer un corpus déjà annoté avec Analec ou Glozz..... | 3 |
| 7.2.3.1 Import de corpus XML-TEI Analec..... | 3 |
| 7.2.3.2 Import de corpus Glozz : à partir de trois fichiers .aa, .aam et .ac..... | 3 |
| 7.2.4 Importer des annotations URS..... | 4 |
| 7.2.5 Annoter des unités interactivement depuis une édition de texte..... | 5 |
| 7.2.5.1 Lancer une session d'annotation..... | 5 |
| 7.2.5.2 Visualiser les unités présentes..... | 5 |
| 7.2.5.3 Créer des unités..... | 5 |
| 7.2.5.4 Éditer les propriétés d'une unité..... | 5 |
| 7.2.5.5 Sélectionner des unités..... | 6 |
| 7.2.5.6 Rechercher des unités par la valeur de leurs propriétés..... | 7 |
| 7.2.5.7 Rectifier les bornes d'une unité..... | 7 |
| 7.2.5.8 Créer des unités à cheval sur deux pages d'édition..... | 8 |
| 7.2.5.9 Supprimer une annotation..... | 8 |
| 7.2.6 Enregistrer les annotations..... | 8 |
| 7.2.7 Annoter, Vérifier, Exploiter, Exporter par commandes..... | 8 |
| 7.2.7.1 Paramètres généraux..... | 9 |
| 7.2.7.1.1 Corpus, Sous-corpus & Partition..... | 9 |
| 7.2.7.1.2 Sélection d'éléments URS..... | 9 |
| 7.2.7.1.2.1 Requêtes URSQL..... | 11 |
| 7.2.7.1.3 Propriétés d'éléments à utiliser..... | 12 |
| 7.2.7.1.4 Autres..... | 12 |
| 7.2.7.2 Commandes d'annotation..... | 13 |
| 7.2.7.2.1 Concordance To Units..... | 13 |
| 7.2.7.2.2 Units Annotate..... | 13 |
| 7.2.7.2.3 Units Delete..... | 14 |
| 7.2.7.2.4 Reset All Annotations..... | 15 |
| 7.2.7.3 Commandes de vérification de la cohérence des annotations..... | 15 |
| 7.2.7.3.1 Check Annotation Structure Values..... | 15 |
| 7.2.7.3.2 Check Duplicates In Schema..... | 15 |
| 7.2.7.4 Commandes d'exploitation d'unités (lister, compter, visualiser...)..... | 15 |
| 7.2.7.4.1 Units Summary..... | 15 |
| 7.2.7.4.2 Units Index..... | 16 |
| 7.2.7.4.3 Units List..... | 17 |
| 7.2.7.4.4 Units Interdistance..... | 18 |
| 7.2.7.4.5 Units Referential Density..... | 19 |
| 7.2.7.4.6 Units Referential Stability..... | 19 |
| 7.2.7.5 Commandes d'exploitation de schémas (lister, compter, visualiser...)..... | 20 |
| 7.2.7.5.1 Schemas Summary..... | 20 |
| 7.2.7.5.2 Schemas Coverage..... | 21 |
| 7.2.7.5.3 Schemas Index..... | 21 |

| | |
|--|----|
| 7.2.7.5.4 Schemas List..... | 22 |
| 7.2.7.5.5 Schema Lengths..... | 23 |
| 7.2.7.5.6 Schemas Progression..... | 23 |
| 7.2.7.5.7 Schema Ratio..... | 24 |
| 7.2.7.6 Commandes d'export..... | 25 |
| 7.2.7.6.1 Export Corpus As..... | 25 |
| 7.2.8 Macros 'Democrat'..... | 26 |
| 7.2.8.1 Vérifier la cohérence des annotations..... | 27 |
| 7.2.8.1.1 Verifications..... | 27 |
| 7.2.8.2 Corriger ou mettre à jour les annotations..... | 27 |
| 7.2.8.2.1 PROPERanaphoriques..... | 27 |
| 7.2.8.2.2 RetoucheComplementDuNom..... | 28 |
| 7.2.8.2.3 Si2Singleton..... | 28 |
| 7.2.8.2.4 SUJ-PARTINF..... | 28 |
| 7.2.8.2.5 TrimPropertyValues..... | 29 |
| 7.2.8.3 Annoter..... | 29 |
| 7.2.8.3.1 Unités..... | 29 |
| 7.2.8.3.1.1 Frpos2Categorie..... | 29 |
| 7.2.8.3.1.2 Fropos2Categorie..... | 30 |
| 7.2.8.3.1.3 AjoutDefinitude..... | 30 |
| 7.2.8.3.1.4 Accessibilite..... | 30 |
| 7.2.8.3.1.5 LongueurDesMentions..... | 30 |
| 7.2.8.3.1.6 PremierMaillon..... | 30 |
| 7.2.8.3.2 Schémas..... | 31 |
| 7.2.8.3.2.1 CreationChaines..... | 31 |
| 7.2.8.3.3 Relations..... | 31 |
| 7.2.8.3.3.1 CreationAnaphores..... | 31 |
| 7.2.8.4 Créer des configurations de corpus..... | 31 |
| 7.2.8.4.1 SubCorpus..... | 31 |
| 7.2.8.4.2 BuildSections..... | 31 |
| 7.2.8.5 Exporter..... | 31 |
| 7.2.8.5.1.1 BuildDemocratCorpusVersion1..... | 31 |
| 7.2.9 Exporter des annotations..... | 32 |
| 7.2.9.1 Dans un corpus binaire TXM..... | 32 |
| 7.2.9.2 Au format XML-TEI URS..... | 32 |
| 7.2.9.3 Au format Glozz..... | 32 |
| 8 Table des matières..... | 33 |