Annotation sémantique et terminologique avec la plateforme SMARTIES

La procédure d'annotation humaine a pour but de distinguer les candidats-termes qui font l'objet d'un emploi terminologique de ceux qui font l'objet d'un emploi en langue courante. L'annotation est réalisée par passages successifs afin d'éviter de réaliser simultanément plusieurs annotations de nature différente. Les couches successives d'annotation répondent chacune à une question élémentaire par deux valeurs : *valide* ou *non valide*. Les quatre étapes de l'annotation évaluent :

- la validité syntagmatique du candidat-terme ;
- l'appartenance du candidat-terme au champ scientifique dont relèvent les textes ;
- l'emploi du terme reconnu en tant que terme dans le contexte précis du document dans lequel il apparaît.

Ainsi, à partir des candidats-termes détectés automatiquement par l'extracteur de TTC-TermSuite (qui réunit les avantages d'Acabit et de Termostat), on sélectionne progressivement des termes relevant du champ scientifique du texte annoté et qui font l'objet d'un emploi terminologique.

Le but de la campagne d'annotation est double : d'une part, il s'agit d'observer les termes d'un champ scientifique en contexte (c'est-à-dire utilisés dans un texte), d'autre part, il s'agit d'enrichir à terme la terminologie de la langue de spécialité du champ scientifique en repérant des termes nouveaux ou des termes non reconnus par l'extracteur automatique du projet (TTC-TermSuite).

Ces deux objectifs ont pour conséquence de ne pas se limiter à ce qui figure dans les ressources pour les étapes d'annotation de vocabulaire de spécialité et d'emploi terminologique. Si on se limite strictement à ce qui est déjà dans les ressources, on perd toute possibilité de faire émerger des termes nouveaux et d'enrichir les ressources. L'enrichissement est donc à ce stade fortement dépendant de l'intuition de l'expert annotateur dans les différentes sciences retenues pour cette expérience.

Le présent guide est structuré en deux parties :

Couche 1 : Annotation syntaxique

(Couche réalisée à l'ATILF)

Cette première couche d'annotation vise à conserver les découpages syntaxiquement corrects et à éliminer les locutions mal découpées, les groupes nominaux complexes incomplets, etc.

A noter que certaines structures recevront un traitement différent en fonction de la discipline.

Critères de sélection :

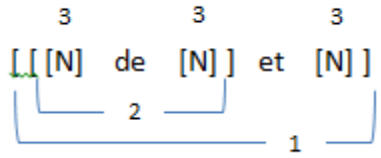
- SN complexes (type : « N de N de N » ou « N de N et N » ou « N de N Adj » ou « N Adj et Adj » ou « N Adj Adj » ou « N Adj de N »)
  - ⇨ Garder le plus petit SN simple correct
  - ⇨ Garder le SN maximal correct
- SN simples (type : « N de N »)
  - ⇨ Garder le SN entier [N de N] et les noms qui le composent [N] de [N]
  - ⇨ Le choix de la frontière à conserver sera fait par l'expert lors de la couche suivante
- SN simples (type : « N Adj »)
  - ⇨ Garder le SN entier [N Adj] et le nom qui le compose [N] Adj
- N isolés
  - ⇨ A garder
- Noms propres isolés
  - ⇨ A garder en Linguistique, Chimie, Sciences de l'info
  - ⇨ A supprimer en Psycho
  - ⇨ En Archéo :
    - o supprimer les noms d'auteurs (ex : « Bordes », « Faivre ») et noms géographiques (c'est-à-dire les entités politiques et administratives de type « France », « Dordogne », « Sarlat »)
    - o garder le reste, c'est-à-dire les noms de sites (ex : « Combe-Grenal »), les noms chronologiques (ex : « Acheuléen », « Moustérien », « Paléolithique »), les noms de technologies (ex : « Quina »), les noms latins de systématique (ex : « Equus caballus piveteaui », « Homo erectus »), les noms topographiques (ex : « sud-ouest », « sud-ouest de la France »)

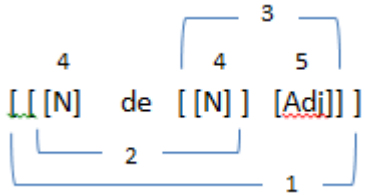
⇨ De manière générale, pour les couples nom/prénom, garder systématiquement l'ensemble nom/prénom ainsi que le nom et supprimer le prénom. Si seul le prénom est retenu, ne rien conserver.

- Adj isolés
  - ⇨ Garder les adjectifs substantivés (type « palatale », « occlusive », « vibrante », « glottale », « sonante ») = adjectifs précédés d'un déterminant
  - ⇨ Supprimer les adjectifs non substantivés partout sauf en chimie

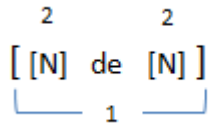
    = adjectifs isolés directement accolés à un nom ou à un syntagme nominal type [N Adj] Adj ou en construction attributive (être/sembler/paraître Adj)
- Verbes isolés
  - ⇨ A supprimer
- Nombres isolés
  - ⇨ A supprimer (ex : [XII]e siècle)

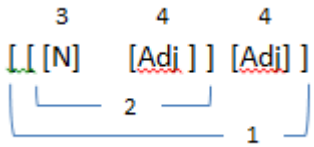
Exemples de groupes nominaux :

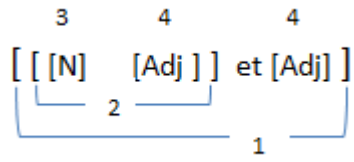
	**1 = on garde** **2 = on garde** **3 = on garde**
N de N et N	
• [• [• [centre] d'• [artistes]] et des • [archivistes]]	
• [• [• [Sciences] de l'• [information]] et de la • [communication]]	

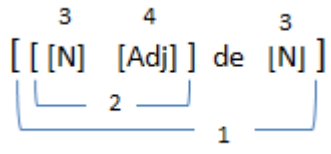
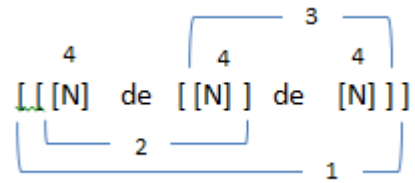
	**1 = on garde** **2 = on supprime quand Adj dépend directement de manière certaine du 2e N (parce que dans ce cas le SN complexe est tronqué et donc syntaxiquement incorrect). Dans les autres cas, on garde** **3 = on garde si morphosyntaxiquement correct** **4 = on garde** **5 = on supprime (sauf en chimie)**
N de N Adj / N à N Adj / N en N Adj	
• [• [• [Hommes] du • [• [paléolithique]] • [supérieur]]]	

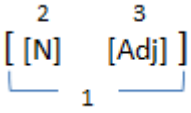
• [• [• [Hommes] du • [• [paléolithique]] • [suivants]]]
• [• [• [Structure] à • [• [attribut]] • [résultatif]]]
• [• [• [syntagme] à • [• [attribut]] • [résultatif]]]
• [• [• [lames] en • [• [silex]] • [tertiaire]]]
• [• [• [lame] en • [• [silex]] • [tertiaire]]]
• [• [• [approvisionnement] en • [• [matières]] • [premières]]]
• [• [•[groupe] de •[• [locuteurs]] • [natifs]]]
• [• [• [série] de • [•[mots]] • [offensants]]]

	**1 = on garde** **2 = on garde**
N prep N	
• [• [coup] de • [burin]]	
• [• [professionnels] de l'• [orientation]]	
• [• [type] de • [profession]]	

	**1 = on garde** **2 = on garde (car on ne sait pas si le 2e Adj est catégorisant)** **3 = on garde** **4 = on supprime (sauf en chimie)**
N Adj Adj	
•[•[• [contraste] • [conceptuel]] • [fort]]	
• [• [• [sens] • [résultatif]] • [commun]]	
• [• [• [contraintes] • [distributionnelles]] • [propres]]	
• [• [• [retouches] • [latérales]] • [inverses]]	
• [• [• [retouches] • [latérales]] • [directes]]	

	**1 = on garde** **2 = on garde** **3 = on garde** **4 = on supprime (sauf en chimie)**
N Adj et Adj / N Adj ou Adj	
• [• [• [différences] • [sémantiques]] et • [pragmatiques]]	
• [• [• [fragments] • [distaux] ou • [mésiaux]]	

	**1 = on garde** **2 = on garde** **3 = on garde** **4 = on supprime (sauf en chimie)**
N Adj de N	
• [• [• [analyse] • [tracéologique]] des • [usures]]	
	**1 = on garde** **2 = on garde** **3 = on garde** **4 = on garde**
N de N de N	
• [• [• [réaction] d'• [• [isomérisation]] du • [N-hexane]]]	

	**1 = on garde** **2 = on garde** **3 = on supprime (sauf en chimie)**
N Adj	
•[• [constructions] • [récentes]]	
•[• [borne] • [initiale]]	
• [• [grammaire] • [constructionnelle]]	
• [• [fouilles] • [récentes]]	
• [• [talon] • [lisse]]	

Exemples de locutions et expressions mal découpées :

Découpage obtenu	*Découpage attendu*
ces •[noms à partir]	ces •[noms] •[à partir de…]
en •[effet l'•[hypothèse]]	•[en effet] l'•[hypothèse]
•[renvoyant à un •[processus] psychologique]	•[•[renvoyant à] •[un processus psychologique]]
prendre également en •[compte des •[noms]]	•[•[prendre également en compte] des •[noms]]
et à mettre en •[évidence qu'il] existe	et •[à mettre en évidence qu'] il existe
en l'•[occurrence] une administration	•[en l'occurrence] une administration
•[prise en •[compte] de •[noms]]	•[•[prise en compte de] •[noms]]

Couche 2 : Annotation du lexique de la discipline

[Couche réalisée à l'INIST]

Cette étape consiste à conserver les candidats termes relevant du lexique propre à la science considérée et à éliminer les candidats termes relevant d'autres sciences ou du lexique transdisciplinaire.

Le lexique à conserver regroupe :

- ***Les candidats termes relevant d'un emploi terminologique.***

Il s'agit d'unités linguistiques qui désignent un concept dans le champ scientifique dont relève le texte.

Exemples : *sociolinguistique, syntaxe, schwa, dictionnaire électronique, variation graphique*

Attention : de manière générale, les frontières des candidats termes qui seront conservés devront représenter les concepts de la discipline tels qu'ils sont invoqués dans le document et non tels qu'ils pourraient exister dans la discipline. Ainsi dans une séquence telle que [[*dictionnaire*] *électronique*] avec deux frontières possibles retenues en couche 1, on privilégiera en couche 2 le candidat terme complexe [*dictionnaire électronique*] et on éliminera [*dictionnaire*] car même s'il a un sens dans la discipline, il n'est pas assez précis pour désigner le concept utilisé dans le document annoté.

- ***Les candidats relevant d'un emploi phraséologique***

Il s'agit d'unités linguistiques propres au discours du domaine considéré. Il peut s'agir d'unités simples que l'on a l'habitude d'utiliser dans les textes du domaine.

Exemples en linguistique : *variante, variation, changement, propriété, unité, type, contraste, norme, système, outil, fonction, usage, forme, construction,*

Il peut également s'agir d'unités complexes apparaissant sous forme de combinaisons de mots ou de termes que l'on a l'habitude d'utiliser ensemble dans les textes du domaine pour caractériser un concept ou pour combiner plusieurs concepts. On parlera aussi dans ce cas de collocations ou de cooccurrences.

Exemples en linguistique : *polarité négative et positive, terme affectif français, compréhension littérale et inférentielle, syllabe initiale*

- ***Les candidats termes provenant d'une autre discipline***

Ces candidats termes doivent être suffisamment intégrés dans le domaine considéré pour faire l'objet d'un emploi terminologique dans celui-ci.

Exemples en linguistique *: apprentissage, enseignement, corpus, communication, récit, base de données, cognition*

- ***Les noms propres***

Ces unités doivent désigner des entités propres au domaine et reconnues par la communauté scientifique.

Exemples en linguistique : *Chomsky, Saussure*

- ***Les sigles***

Ces unités doivent être reconnues par la communauté scientifique.

Exemples en linguistique : *TAL*

Le lexique à éliminer regroupe :

- ***Les candidats termes relevant d'autres sciences***

Ces candidats termes sont éliminés lorsqu'ils ne sont pas intégrés dans le domaine scientifique des textes.

Exemples en linguistique : *France, église, statue, carte conceptuelle, élève*

- ***Les candidats termes relevant du lexique transdisciplinaire***

Il s'agit d'un lexique propre aux écrits scientifiques qui traverse les disciplines et qui sert à la description et à la présentation de l'activité scientifique, des procédures scientifiques, du raisonnement, de la filiation et du positionnement scientifique.

Exemples : *étude, hypothèse, méthodologie, application, processus, stratégie*

Couche 3 : Annotation des emplois terminologiques

[Couche réalisée à l'INIST]

Cette étape consiste à conserver les candidats reconnus en tant que termes dans le contexte précis du document dans lequel ils apparaissent et à éliminer les candidats termes relevant du lexique et de la phraséologie de la discipline.

Les unités linguistiques à conserver regroupent :

- ***Les candidats faisant l'objet d'un emploi terminologique***

Il s'agit des candidats termes qui désignent un concept. Ceux-ci doivent pouvoir faire l'objet d'une définition et doivent se présenter dans la forme qu'ils pourraient avoir dans une terminologie du domaine.

Exemples en linguistique : *sociolinguistique, syntaxe, schwa, dictionnaire électronique, variation graphique*

A noter que les synonymes sont acceptés ainsi que les variations en nombre.

- ***Les unitermes dont le sens ne change pas lorsqu'ils sont pris de manière isolée***

Exemples : [*syllabe]* dans *[[syllabe] initiale]* car l'adjectif « *initiale* » n'a pas une fonction catégorisante.

- ***Les termes provenant d'une autre discipline***

Ces termes doivent faire l'objet d'un emploi terminologique stabilisé et bien ancré dans le domaine considéré et dans le contexte donné.

Exemples en linguistique *: apprentissage, enseignement (*dans des textes de psycholinguistique*)*

- ***Les noms propres***

Ces unités doivent désigner des entités propres au domaine et reconnues par la communauté scientifique.

Exemples : *Chomsky, Saussure*

- ***Les sigles***

Uniquement les sigles reconnus par la communauté scientifique.

Exemples en linguistique : *TAL*

Unités linguistiques à supprimer :

- ***Les unités phraséologiques***

Il s'agit des unités linguistiques propres au discours du domaine considéré, qu'elles soient simples ou complexes.

Les unités phraséologiques simples sont en effet trop génériques pour apparaître dans une terminologie du domaine. Exemples en linguistique : *propriété, unité, type, système, outil*

Les unités phraséologiques complexes quant à elles ne représentent pas un nouveau concept, mais une caractérisation d'un concept existant ou une combinaison de concepts. Exemples en linguistique : *polarité négative et positive, terme affectif français, syllabe initiale*

Guide d'utilisation de l'interface SMARTIES

L'interface d'annotation permet de se connecter, de choisir le texte sur lequel travailler, d'effectuer les annotations, d'en sauvegarder le résultat puis de préparer le texte pour la couche d'annotation suivante.

Se connecter

Dans un navigateur web, se rendre à l'adresse : https://apps.atilf.fr/smarties/

Sur la page d'accueil, se connecter en cliquant sur "Login" et en donnant le nom d'utilisateur et le mot de passe qui vous a été attribué par E. Jacquey, B. Husson ou S. Barreaux.

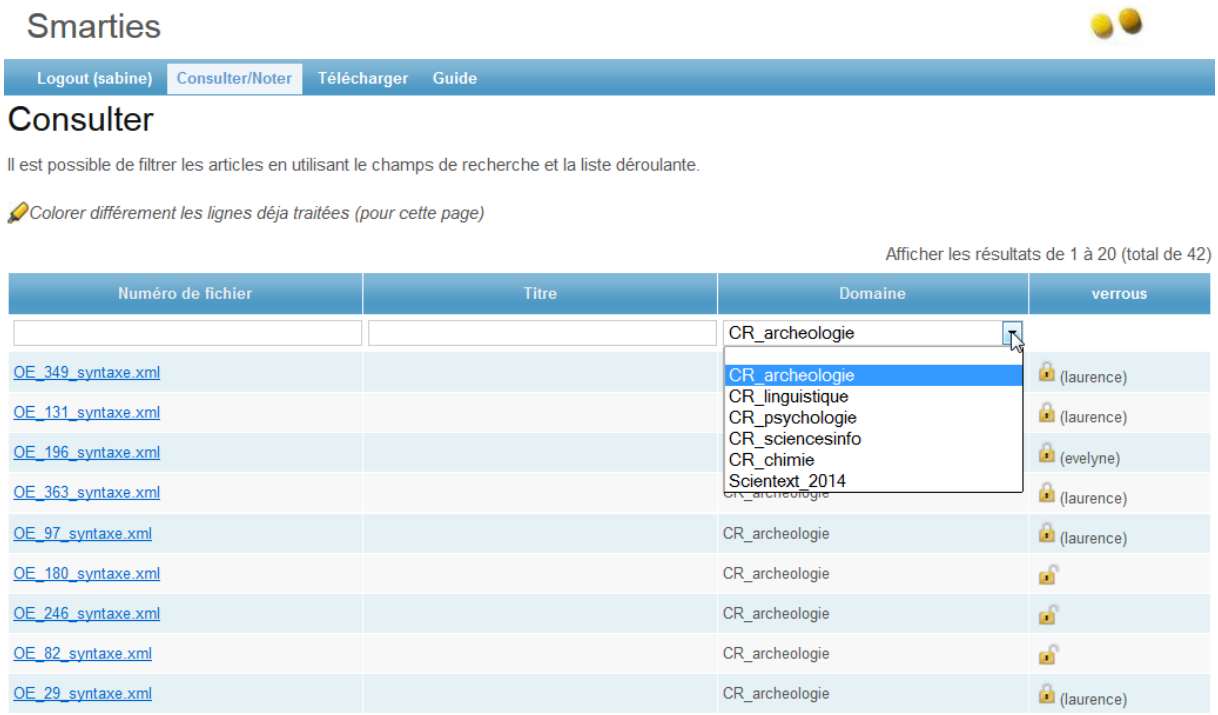
Après cette première étape, vous revenez à la page d'accueil dans laquelle il faut cliquer sur "Consulter/Noter" pour démarrer une annotation.

Choisir un texte sur lequel travailler

Après avoir cliqué sur "Consulter/Noter", vous arrivez sur la page qui affiche la totalité des textes de la base de données qui contient tous les fichiers, chargés, annotés, enregistrés via l'interface.

Pour faciliter le repérage du texte sur lequel vous voulez travailler (commencer ou poursuivre), vous avez deux manières de limiter le nombre total de textes qui s'affichent :

- en utilisant la sélection d'un domaine : dans la figure ci-dessous, si vous cliquez sur le champ "domaine", l'interface vous propose tous les domaines enregistrés jusque-là. Les domaines correspondant aux expériences en cours sont les suivants : Archéologie, Linguistique, Psychologie, Sciences de l'information, Chimie et Scientext 2014.



• la seconde façon de réduire le nombre de fichiers qui s'affichent est de donner son numéro dans le champ "Numéro de fichier". Comme on le voit ci-dessous, on peut accéder directement aux versions existantes du fichier "article_97" en tapant "97".

La figure ci-dessus montre 3 versions du même fichier : les fichiers suffixés par "syntaxe", "disciplinaire" et "terminologique" correspondent au résultat de chacune des 3 couches d'annotation expliquée dans la première partie de ce guide.

En utilisant le champ « Numéro de fichier », on peut aussi indiquer la couche à laquelle on veut annoter (exemple : « _disciplinaire ») pour avoir la liste de tous les textes à annoter en couche 2.

A noter que, pour repérer facilement dans la liste les fichiers terminés et les fichiers en cours de travail ou non commencés, il suffit de cliquer sur la phrase « Colorer différemment les lignes déjà traitées ». Comme l'illustre l'image ci-dessous, cette action permettra de distinguer les fichiers pour lesquels le fichier de la couche suivante a été généré.

Enfin, pour boucler le choix d'un texte et l'afficher, il suffit de cliquer sur son nom.

Effectuer les annotations

Une fois un texte choisi et affiché, on peut procéder à son annotation, couche par couche.

Avant de commencer, il faut verrouiller le texte sur lequel on va travailler de façon à éviter qu'une autre personne ne modifie par mégarde les annotations déjà réalisées. Pour cela, dès l'affichage, il faut cliquer sur le petit verrou en haut à gauche afin de "verrouiller la ressource".

Une fois ceci fait, le message "*Vous n'avez pas les autorisations suffisantes pour enregistrer des modifications sur res_scientext104.xml* " indiqué en haut du texte disparaît et le nom de l'utilisateur apparaît entre parenthèses à côté du verrou qui est maintenant fermé. (evelyne)

Si vous commencez à annoter avant d'avoir verrouillé le texte, les annotations réalisées seront perdues.

Pour l'annotation proprement dite, il s'agit de mettre en rouge les pastilles qui correspondent à des candidats non valides selon les critères de la couche d'annotation dans laquelle on se trouve. Pour rappel, au début de l'annotation, tous les candidats sont précédés de pastilles vertes, c'est-à-dire qu'ils sont considérés comme valides par défaut. Pour faire passer une

pastille du vert au rouge, il suffit de cliquer dessus. Pour faire repasser une pastille du rouge au vert, cliquer à nouveau dessus.

Dans la figure ci-dessous, on peut voir sur les premières lignes, l'application des critères de la couche 1 (syntaxe). Sur cette figure, on observe aussi que l'on peut connaître les limites du groupe dont on doit décider s'il est valide ou non en observant le texte colorié en bleu au moment où on passe sur un point, vert par défaut.

Une fois l'ensemble du texte annoté, on va sauvegarder ce résultat puis préparer le texte pour la couche d'annotation suivante.

Sauvegarde des résultats intermédiaires et préparation des textes pour les couches d'annotation suivantes

En cours d'annotation, chaque décision de validité et d'invalidité est enregistrée automatiquement et instantanément. On peut donc interrompre le travail en cours, quitter la plateforme et fermer le navigateur en toute sécurité sans perdre son travail.

A l'issue de l'annotation, pour sauvegarder les décisions de validité et d'invalidité qui ont été prises et pour préparer le texte pour la suite, une seule opération suffit : cliquer sur le bouton "Créer...".

Après avoir effectué cette opération, un nouveau fichier est créé : il porte le nom indiqué dans le bouton sur lequel on vient de cliquer, il correspond à l'état final de la couche d'annotation que l'on vient de finir et à l'état initial de la couche d'annotation que l'on va aborder ensuite.

Ainsi, dans l'exemple ci-dessus, res_scientext104_syntaxe.xml correspond au résultat de la première couche d'annotation mais sera aussi le fichier que l'on va afficher lorsqu'on va démarrer la seconde couche d'annotation (repérage des candidats termes relevant lexique disciplinaire).

Pour mémo, voici un tableau récapitulatif des noms de fichier à l'entrée et à la sortie de chaque couche :

Couche	Entrée	Sortie
1	Domaine_n°.xml	Domaine_n°_**syntaxe**.xml
2	Domaine_n°_**syntaxe**.xml	Domaine_n°_ **disciplinaire**.xml
3	Domaine_n°_**disciplinaire**.xml	Domaine_n°_**terminologique**.xml

A noter qu'il n'est pas prévu de retour facile sur une couche d'annotation. Il est donc recommandé de rester très vigilant au moment de la création du document pour la couche suivante et d'être sûr de ses décisions. En cas d'erreur importante constatée sur une couche précédente, il faut contacter l'administrateur de l'interface.

Il vous est demandé également de noter le temps passé sur chaque couche pour l'ensemble du corpus de votre discipline. Pour cela, vous pouvez utiliser Time Tracker (http://www.formalassembly.com/time-tracker).

Pour toute opération sur l'interface, il faut contacter ses administrateurs :

Benjamin Husson, Etienne Petitjean, Evelyne Jacquey ou Sabine Barreaux.

Bon courage à tou(te)s