Station Sensunique: une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non terminologiques (orientée Langues Contrôlées)

Izabella Thomas[1], Blandine Plaisantin Alecu[2], Bérenger Germain[3], Marie-Laure Betbeder[4]
[1]Centre L. Tesnière, Université de Franche-Comté
[2]Prolipsia, France
[3]Share and Move Solutions, France
[4]Institut Femto-ST, Université de Franche-Comté
izabella.thomas@univ-fcomte.fr, blandine.alecu@prolipsia.com,
berenger.germain@shareandmove.fr, marie-laure.betbeder@univ-fcomte.fr

Résumé

Dans cet article, nous présentons le fonctionnement et les services proposés par la Station Sensunique, une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et du lexique d'une Langue Contrôlée. La Station prend en charge et facilite l'ensemble de ce processus, à travers une analyse automatique du corpus, puis la possibilité d'approfondir l'analyse manuellement, pour aboutir à la création des fiches terminologiques/lexicales et des ressources exportables. La Station est un outil prêt à l'emploi, ergonomique, facile à prendre en main et à exploiter à travers ses différentes interfaces. Elle allie les avantages d'une analyse automatique (rapidité, coût) avec l'exactitude et la fiabilité d'une analyse humaine.

Mots-clés: lexique; langue contrôlée; ressource terminologique; extraction des termes; acquisition des termes; plateforme terminologique

1 Description générale

La Station Sensunique est une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non terminologiques. Elle a été conçue à l'Université de Franche-Comté durant le projet ANR-EMMA-2010-039 intitulé Sensunique[1] (2011-2012) dans l'objectif d'accélérer le processus d'établissement du lexique d'un domaine ou d'une Langue Contrôlée (LC). Elle prend en entrée un corpus de textes et produit en sortie des diverses ressources (dictionnaires, lexiques, glossaires) enrichies de multiples informations linguistiques. Elle s'appuie sur une analyse automatisée de corpus, dont les résultats sont la base d'une phase de validation manuelle effectuée par un analyste, puis par un expert d'un domaine. La spécificité de cette station par rapport à d'autres

1 http://tesniere.univ-fcomte.fr/sensunique.html [08/04/2014].

plateformes de travail terminologique (HyperTerm[2], Terminae[3], Terminus[4]), repose (Thomas et al. 2014) :

- d'une part, sur les choix méthodologiques sur lesquels elle est fondée : la collaboration de plusieurs outils TAL (Plaisantin Alecu et al. 2012), l'interrogation automatique des ressources terminologiques existantes, l'intégration et l'interrogation des ressources terminologiques ou lexicales propres; ceci en vue de faciliter le travail de l'analyste en lui proposant une liste d'Unités Lexicales Candidates (ULC) pondérées et enrichies de multiples informations linguistiques acquises automatiquement ;
- d'autre part, sur les objectifs spécifiques desquels elle découle, dont notamment le recensement du Lexique d'une Langue Contrôlée (LLC), définie comme une *Langue Contrôlée sur mesure* (Renahy et al. 2009, 2011).

La spécificité d'un tel lexique est qu'il se doit d'être exhaustif : toutes les unités nécessaires lors de l'écriture de documents, qu'elles soient ou non terminologiques, doivent être encodées (pour être utilisables) dans le dictionnaire d'une LC. De plus, cette contrainte d'exhaustivité du niveau lexical d'une LC implique de distinguer au moins deux types de dictionnaires : un dictionnaire du lexique d'une LC et un dictionnaire des structures lexicales (Thomas et al. 2014), chacun pouvant être soit terminologique soit général. Une autre contrainte liée à la conception de LLC provient des principes d'une LC : de non-ambigüité (à une Unité Lexicale (UL) ne correspond qu'un sens) et, inversement, de non-redondance (à un sens correspond une et une seule UL). Ceci présuppose la gestion de la synonymie, et plus généralement la gestion des relations entre plusieurs unités lexicales (telles que homonymie, dérivation, collocations etc.).

La station est orientée analyste – utilisateur. Tous les résultats (ULC ou informations associées) sont des propositions que l'utilisateur peut modifié (ajouter, modifier, compléter, valider ou invalider). Il est assisté dans ce processus par un ensemble de fonctionnalités d'exploration des résultats, à savoir : visualisation des ULC sous plusieurs modes, tris et filtres sur la liste des ULC, projection des ULC sur le corpus d'origine, regroupement de différentes ULC, recherches sur le corpus à l'aide d'un concordancier avancé, etc.

L'utilisation de la Station se fait, sans contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisées par la Station. Enfin, la Station Sensunique est dotée d'une interface utilisateur facile à manier et à explorer.

2 http://www.tedopres.com/hyperterm-terminology-management [08/04/2014].
3 http://lipn.univ-paris13.fr/terminae/index.php/Main_Page [08/04/2014]
4 http://terminus.iula.upf.edu//cgi-bin/terminus2.0/terminus.pl [08/04/2014].

2 Architecture et services

La station Sensunique fonctionne de façon modulaire, chaque module proposant à l'utilisateur plusieurs services. Les modules sont organisés pour correspondre au processus d'acquisition de ressources, divisé en plusieurs étapes (représenté par la Figure 1).

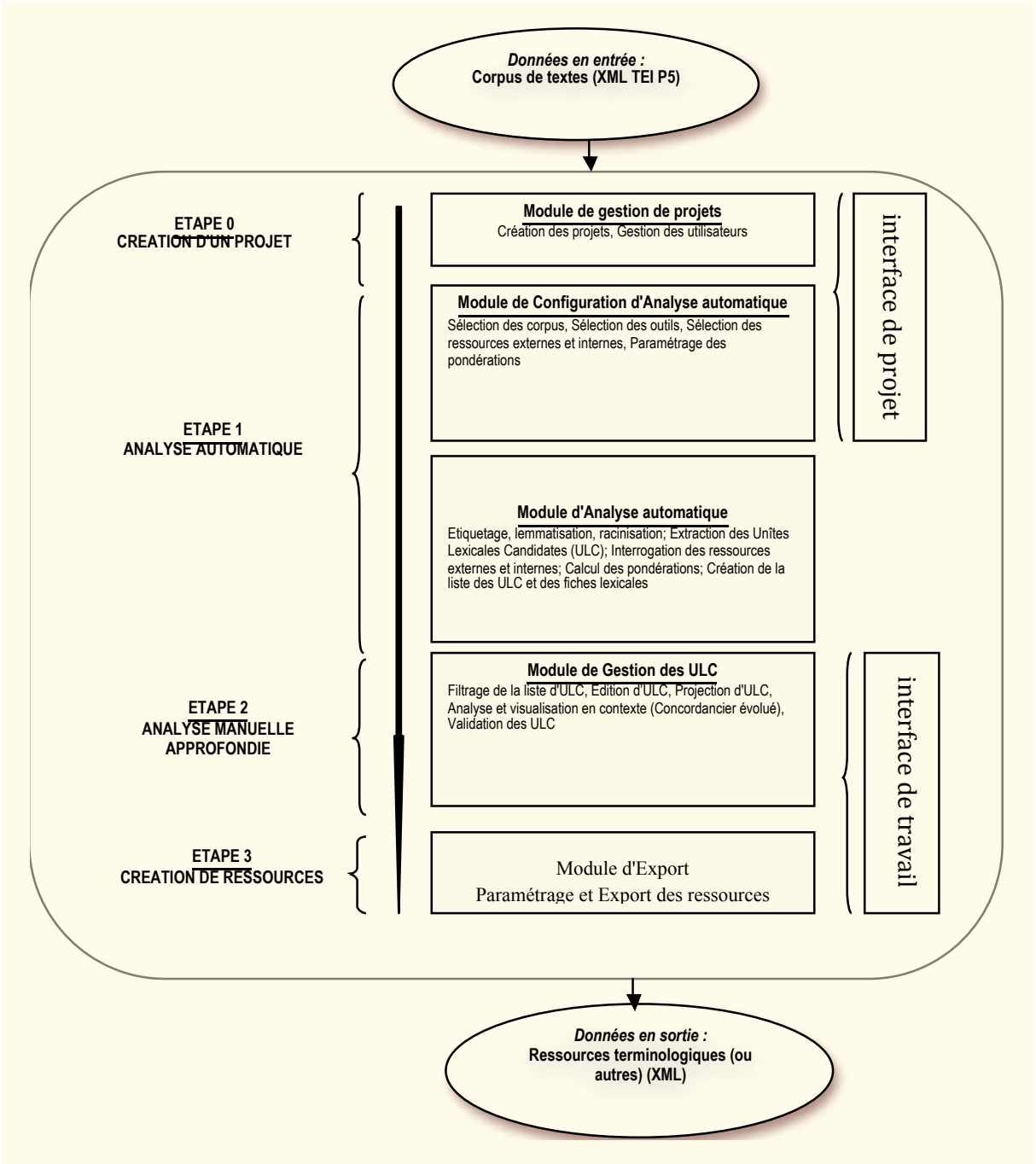

Figure 1: Schéma de la Station Sensunique.

L'utilisateur peut interagir avec la Station à partir de deux interfaces : (1) *l'interface de projet*, qui sert à définir et paramétrer un projet en termes d'utilisateurs, de corpus, d'outils et de ressources utilisés pour l'analyse automatique et (2) *l'interface de travail*, qui permet d'explorer les résultats d'analyse automatique en vue de l'établissement d'une ressource finale. Nous présentons les diverses fonctionnalités de la Station en suivant le processus chronologique d'un utilisateur souhaitant implémenter un nouveau projet.

2.1 Etape 0: Création d'un projet

La création d'un projet commence par l'établissement d'un groupe de travail, c'est-à-dire par la déclaration d'un ou plusieurs utilisateurs ayant le droit de travailler sur le projet. En effet, la Station est collaborative : elle permet à plusieurs utilisateurs d'interagir sur la même tâche. Elle assure aussi la traçabilité de toute modification (correction, modification ou complétion des données) grâce à une étiquette portant le nom de l'utilisateur concerné. Ces étiquettes de traçabilité permettent également de distinguer les données obtenues de façon automatique des données créées ou modifiées par un utilisateur.

Un groupe de travail peut créer plusieurs projets (Figure 2) ; chaque projet, en plus du nom et de sa date de création, est caractérisé par son domaine et le public auquel il est destiné. Un projet ne peut contenir qu'un corpus pour chaque type de corpus permis : Corpus d'Analyse, Corpus Support, Corpus Contrastif (Thomas et al., 2014). Les corpus doivent être chargés dans la Station au format XML TEI P5[5]; la conversion de tout document vers ce format doit être faite au préalable en utilisant, par exemple Oxgarage[6], un convertisseur automatique de format de documents en ligne.

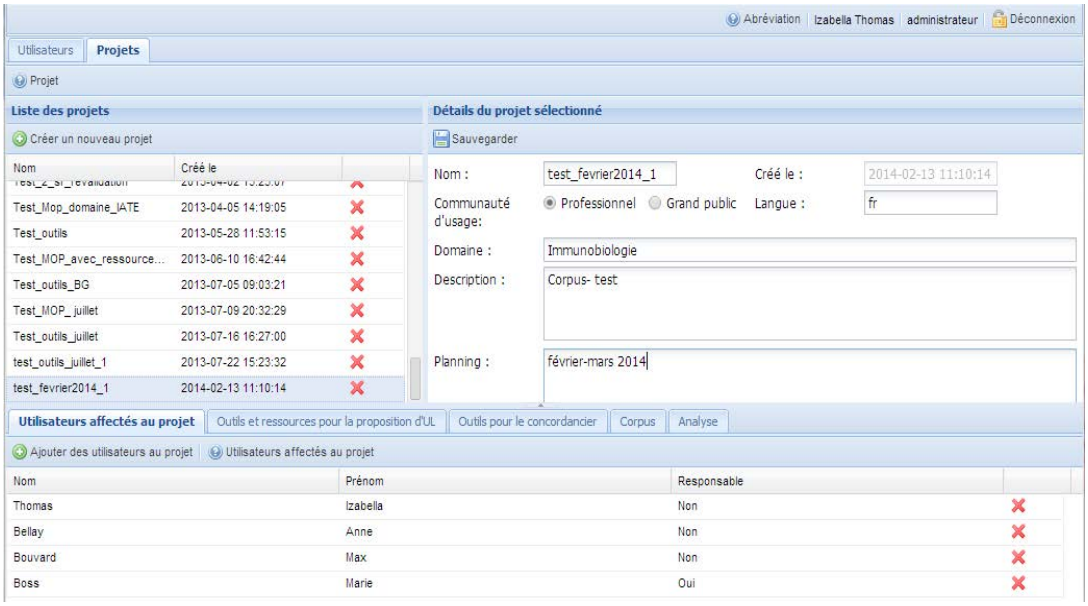

Figure 2: Déclaration d'un projet (capture d'écran).

5 http://www.tei-c.org/Guidelines/Customization/Lite/teiu5_fr.html [04/04/2014].
6 http://oxgarage.oucs.ox.ac.uk:8080/ege-webclient [accédé le 08/04/2014].

2.2 Etape 1: Analyse automatique

Cette étape est composée de deux sous-étapes : (1) la configuration de l'analyse automatique par l'utilisateur et (2) l'analyse automatique.

La Station Sensunique est hautement paramétrable (Thomas et al. 2014), dans l'objectif d'assurer l'adéquation de l'analyse avec la ressource à construire. Les paramétrages se font à partir de l'interface de projet (Figure 3): en fonction de son objectif, l'utilisateur peut choisir les chaînes d'outils et les ressources externes à interroger, paramétrer l'algorithme de pondération et incorporer des ressources internes (moyennant leur mise en format appropriée). Le choix de ressources internes n'est pas restreint, ce qui assure à la Station son caractère évolutif.

La qualité de l'analyse et le nombre d'informations recueillies par la Station Sensunique à partir de diverses ressources dépend du paramétrage effectué par l'utilisateur.

Actuellement, les outils et les ressources intégrés à la Station Sensunique sont les suivants :

- étiqueteurs morphosyntaxiques : TreeTagger[7] et Brill[8] (Brill 1992) ;
- analyseur flexionnel du français : Flemm v2 et v3 (Namer, 2000) ;
- extracteurs de termes : Acabit (Daille 1994), TermoStat (Drouin 2003) et YaTeA (Aubin et al. 2006) ;
- racinisateur Lingua:: Stem[9] ;
- TermSciences[10], portail terminologique multidisciplinaire développé par CNRS-INIST ;
- IATE[11], base de données terminologique de l'Union Européenne.

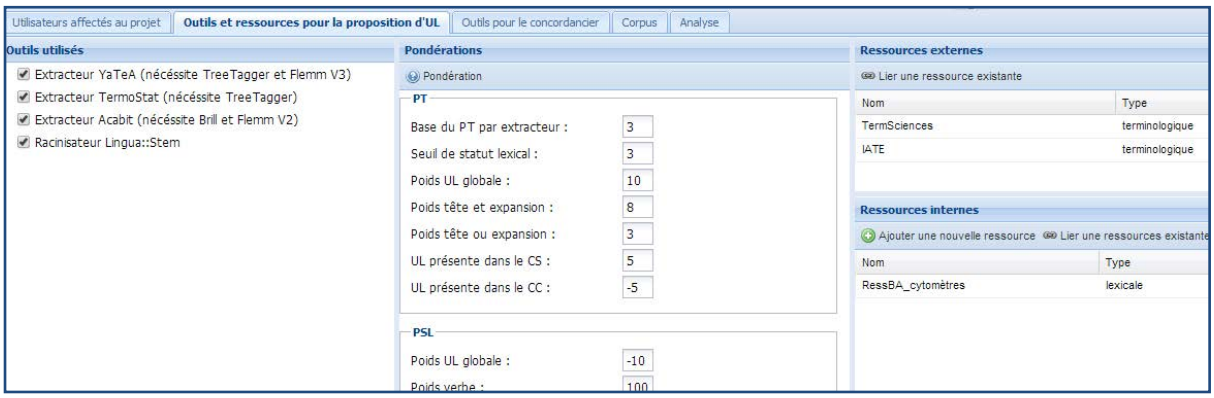

Figure 3: Sélection des outils et des ressources (capture d'écran).

En ce qui concerne la durée de l'analyse automatique, elle dépend de la taille du corpus et du paramétrage utilisateur. Pour un corpus de 50 fichiers représentant un volume de 507 Ko, soumis aux outils TreeTagger et Flemm v3, jugés représentatifs du comportement global des outils lors d'une analyse, le temps d'exécution s'élève à 7,735 s. Il n'augmente que modérément avec l'augmentation du volume

7 http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schmid/Tagger-Licence [08/04/2014].
8 Avec le lexique et le fichier de règles fournis par l'ATILF-CNRS, de Nancy.
9 http://search.cpan.org/~sdp/Lingua-Stem-Fr0.02/lib/Lingua/Stem/Fr.pm [04/12/2011].
10 http://www.termsciences.fr/ [08/04/2014].
11 http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load [08/04/2014].

d'informations à traiter (3,486 s pour un 1 corpus de 20 fichiers représentant un volume de 214 Ko). Par contre, l'interrogation des ressources externes (par web services) peut rallonger considérablement le temps d'exécution (jusqu'à plusieurs heures).

2.3 Etape 2: Analyse manuelle approfondie

Les résultats de l'analyse automatique sont affichés dans l'interface de travail. Cette interface est divisée en 4 espaces (Figure 4) que l'on peut repositionner, redimensionner, afficher ou cacher.

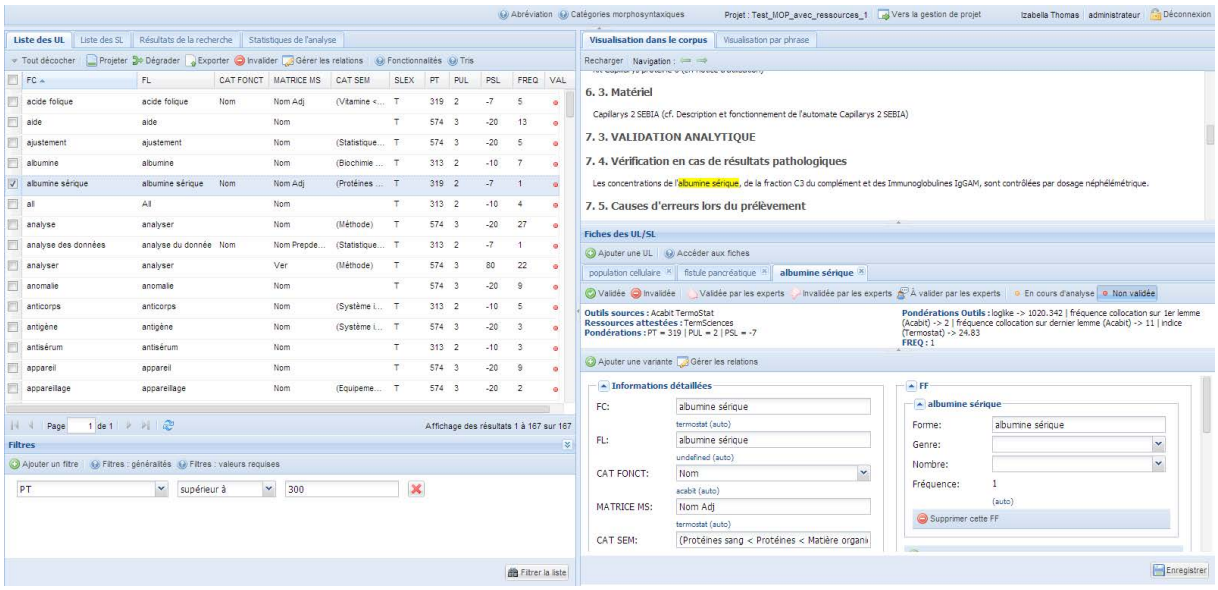

Figure 4: Interface de travail, vue globale (capture d'écran).

Dans l'espace 1, l'analyste visualise la liste des ULC assorties d'informations servant aux tris (par simple clic sur la colonne correspondante) et filtrages des résultats. L'espace 2 permet de filtrer la liste des ULC à l'aide de 21 paramètres, telles que la fréquence, l'extracteur-source ou la classe sémantique. Les filtres sont cumulatifs : il est possible de filtrer la liste selon plusieurs paramètres simultanément ; par exemple, les ULC de matrice morphosyntaxique Nom Adj, proposées par TermoStat, avec au moins 20 occurrences dans le corpus. L'espace 3 sert à visualiser les ULC dans leur contexte initial, en projetant une ou plusieurs ULC sélectionnée(s) dans la liste, en corpus ou par phrase (cf. exemple de «*albumine sérique*» sur la Figure 4). L'espace 4 sert à afficher les fiches lexicales des ULC sélectionnées dans la liste. Une fiche lexicale comporte l'ensemble d'informations concernant l'ULC, c'est-à-dire une description complète de la forme canonique d'une ULC avec des spécifications sur ses formes fléchies. L'analyste peut modifier, ajouter, valider ou enlever les informations.

Chaque ULC est aussi assortie d'une *fiche de relations* qui permet de visualiser et/ou de définir un réseau de relations qu'elle entretient avec d'autres formes recensées. Il s'agit de :

- relations morphologiques (recensement de formes fléchies (FF) d'une ULC, recensement de formes en relation de dérivation avec une (partie de) ULC (cf. UL dérivées sur la Figure 5) ;

- relations lexico-syntaxiques (cf. UL incluses, composées et associées sur la Figure 5) ;
- relations lexico- sémantiques (cf. UL homonymes, UL variantes sur la Figure 5).

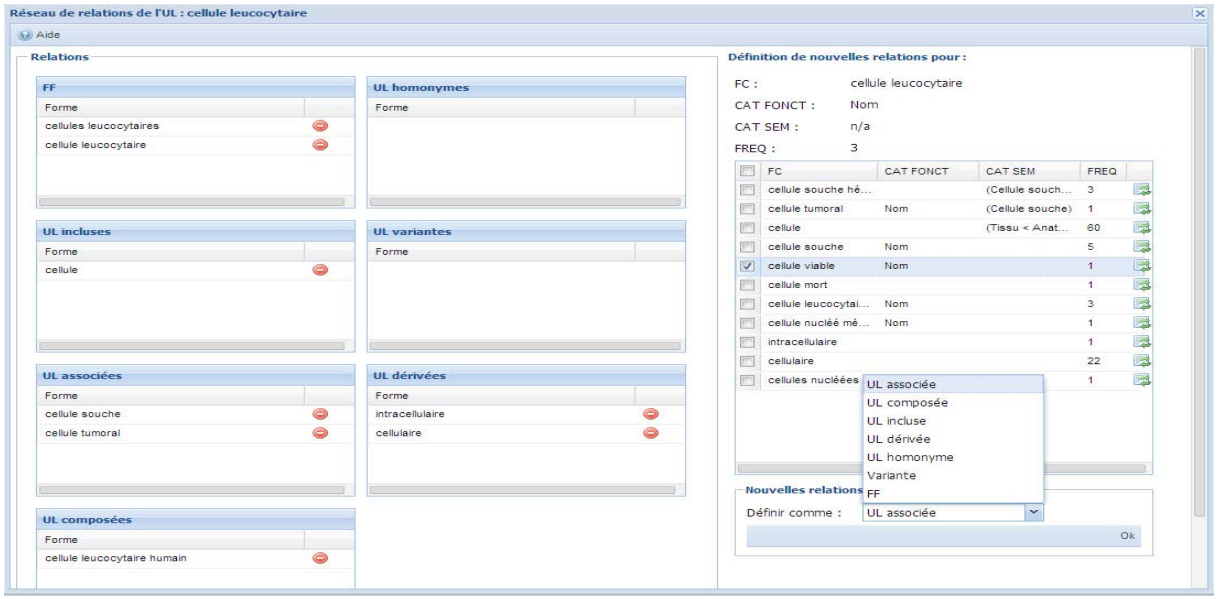

Figure 5: Exemple d'une fiche de relations pour UL *cellule leucocytaire* (capture d'écran).

Certaines relations entre les ULC sont proposées automatiquement, soit par les outils/ ressources, soit à partir de calculs effectués par la station. L'analyste peut les valider/invalider, mais aussi établir des nouvelles relations (cf. Figure 5, «*cellule viable*» sélectionnée dans la liste de gauche est définie comme une UL associée à «*cellule leucocytaire*»).

Afin d'approfondir l'analyse en explorant le corpus, la Station Sensunique propose un concordancier évolué, qui permet de visualiser les occurrences d'une forme en contexte, dans le corpus ou dans les phrases isolées (Figure 6). La recherche s'effectue soit directement à partir des informations saisies par l'utilisateur, soit en demandant à la Station de calculer les informations linguistiquement plus complexes (lemmes, racines ou catégories morphosyntaxiques) concernant les formes à rechercher. Le concordancier est dit 'évolué' au sens où il permet différents types de recherche, allant d'une simple recherche sur une chaîne de caractères jusqu'à une recherche impliquant la combinaison de différents critères linguistiques (lemmes, racines, catégories morphosyntaxiques d'une ou plusieurs unités lexicales). Combiner des critères appartenant à différents niveaux d'analyse permet d'imposer des contraintes plus ou moins fortes sur les motifs recherchés, et ainsi cibler (ou, au contraire, élargir) le champ des résultats. Par exemple, la recherche '*[l]cellule [e]Adj*' (exprimée sous forme d'expressions régulières Sensunique) permet de cibler les groupes composés d'une des formes fléchies du mot «*cellule*» suivie d'une forme à fonction adjectivale (ex. «*cellules nucléées*», «*cellules totales*», «*cellules mortes*» etc.) (cf. Figure 5).

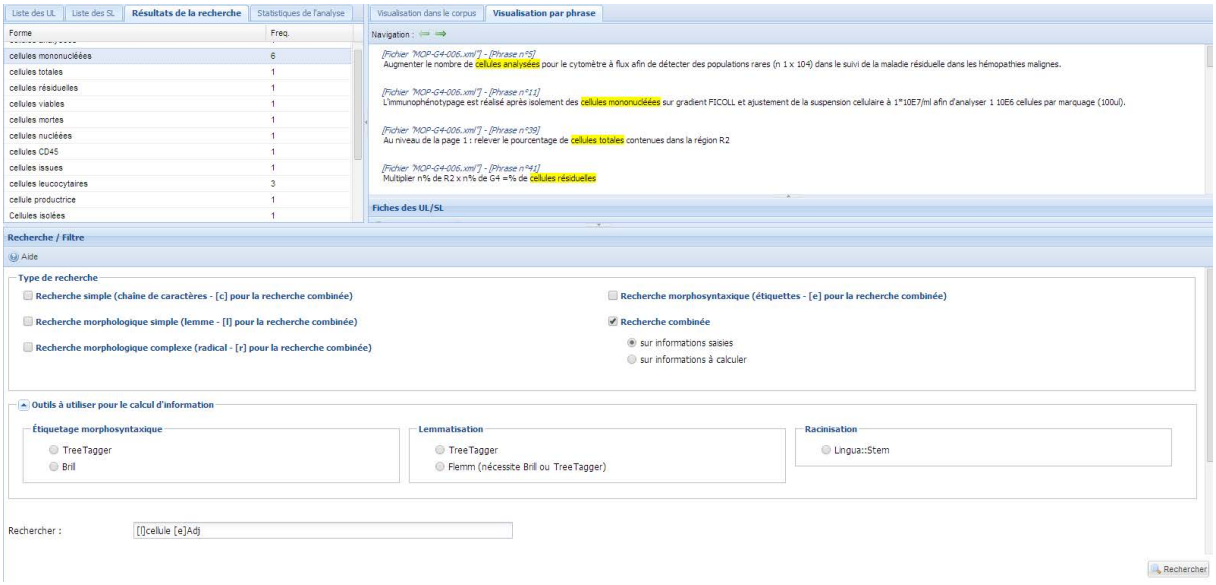

Figure 6: Concordancier évolué (capture d'écran).

2.4 Etape 3 : Création de ressources

A tout moment de son travail, l'utilisateur peut exporter les données recensées dans la station afin de les exploiter dans d'autres applications, les valider par un expert-métier ou simplement, créer une ressource terminologique finale. Les données à exporter, au format XML, peuvent être sélectionnées et restreintes à certaines valeurs grâce à un système de filtres cumulatifs (Figure 7).

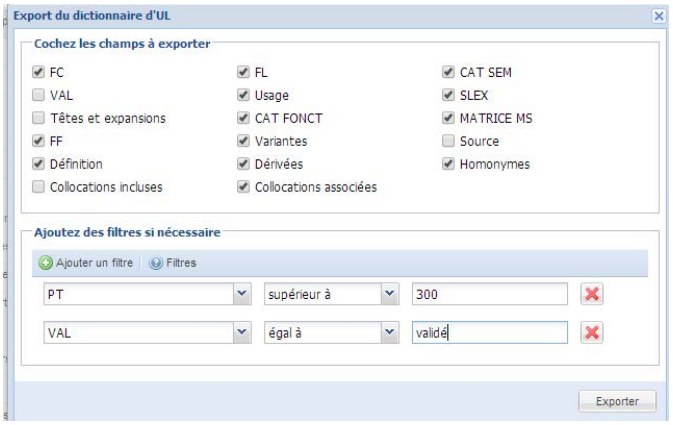

Figure 7: Paramétrage de l'export (capture d'écran).

3 Conclusion

Les fondements méthodologiques et l'architecture logicielle de la Station Sensunique lui permettent de dépasser son objectif initial (établissement du lexique d'une LC) et lui donne le potentiel d'être un

outil générique assistant l'établissement de diverses ressources terminologiques : glossaires, dictionnaires, bases de données, thesaurus, index, termino-ontologies etc., aussi bien pour une consultation directe que comme entrées pour d'autres applications en TAL (recherche et extraction d'information, systèmes d'indexation, acquisition et représentation des connaissances etc.). La facilité d'utilisation de la Station Sensunique nous semble un véritable avantage : c'est un outil prêt à l'emploi, ergonomique, facile à prendre en main et à exploiter. Elle allie les avantages d'une analyse automatique (rapidité, coût) et l'exactitude et la fiabilité d'une analyse humaine. L'utilisation de la Station se fait sans aucune contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisées par la Station.

Nous projetons d'intégrer de nouvelles fonctionnalités à la Station Sensunique, concernant le traitement du contenu sémantique des textes, principalement l'amélioration de la détection des relations sémantiques et conceptuelles entre les unités lexicales. Une autre direction de recherche inclut la construction d'ontologies de domaine.

4 Références bibliographiques

Aubin S. et Hamon, T. (2006). Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing*, 5th International Conference on NLP (FinTAL'2006), Springer, p. 380-387.

Brill E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing* (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155.

Daille B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Workshop at the 32nd Annual Meeting of the ACL (ACL'94), Las Cruces, New Mexico, USA.

Drouin P. (2003). Term Extraction Using non-Technical Corpora as Point of Leverage. In *Terminology*, vol.9, n°1, John Benjamins Publishing Company: Amsterdam/Philadelphia, p. 99-115.

Namer, F. (2000). FLEMM: un analyseur flexionnel du français à base de règles. In *Traitement Automatique des Langues*; vol. 41/2, p. 523-547.

Plaisantin Alecu B., Thomas I., Renahy J. (2012). *La « multi-extraction » comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques*, In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012,* volume 2 : TALN, ATALA/AFCP, pp.511 518, http://www.aclweb.org/anthology/F/F12/F12-2047.

Renahy J., Devitre D., Thomas I., Dziadkiewicz A. (2009). Controlled language norms for the redaction of security protocols: finding the median between system needs and user acceptability, in *Proceedings of the 11th International Symposium on Social Communication*, Santiago de Cuba, Cuba, 19-23 January 2009, pp. 289-293.

Renahy J., Thomas I., Chippeaux G., Germain B., Petiaux X., Rath B., De Grivel V., Cardey S., Vuitton DA. (2011). La langue contrôlée et l'informatisation de son utilisation au service de la qualité des textes médicaux et de la sécurité dans le domaine de la santé, In P. Staccini, A. Harmel, S. Darmoni, R. Gouider, *Systèmes d'information pour l'amélioration de la qualité en santé*, Comptes rendus des quatorzièmes Journées francophones d'informatique médicale (JFIM'2011), Tunis, 23-24 septembre 2011, Springer-Verlag.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of international conference on new methods in language processing (Vol. 12, pp. 44-49).

Thomas I., Plaisantin Alecu B., Germain B., Betbeder M.L. (2014). Station Sensunique: Architecture générale d'une plateforme web paramétrable, modulaire et évolutive d'acquisition assistée de ressources. In *Preceedings of Euralex 2014*, EURAC, Institute for Specialised Communication and Multilingualism.