Station Sensunique: Architecture générale d'une plateforme web paramétrable, modulaire et évolutive d'acquisition assistée de ressources

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 décrivons l'architecture générale d'une plateforme web paramétrable, modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non-terminologiques : la Station Sensunique. Conçue dans l'objectif initial de faciliter et d'accélérer le processus de constitution du lexique d'une Langue Contrôlée (LC), son champ d'application peut être élargi à l'acquisition de tout type de ressources termino-ontologiques. La Station s'articule autour de deux points de vue du processus d'acquisition de ressources : (1) chronologique (centré processus) : import des textes d'entrées, analyse automatique, analyse manuelle approfondie, validation, et enfin export; (2) ergonomique (centré utilisateur-analyste) : mise en adéquation de l'analyse selon le corpus et l'application visée, visualisation et gestion des unités lexicales candidates (ULC), recherches complexes en corpus ou dans la liste d'ULC, modification ou enrichissement des descriptions ou relations des ULC, validation progressive, demande de validation par l'expert-métier, etc. La Station dispose de deux interfaces utilisateurs faciles à manipuler ; son utilisation se fait sans aucune contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisées.

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

1 Introduction

La Station Sensunique est une plateforme web paramétrable, modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non-terminologiques créée à l'Université de Franche-Comté dans le cadre du projet ANR Sensunique[1]. Conçue dans l'objectif initial d'assister le

1 Projet ANR -EMMA-2010-039 (2010-12), http://tesniere.univ-fcomte.fr/sensunique.html [08/04/2014].

processus de constitution du lexique d'une Langue Contrôlée (LC), telle que définie par (Renahy et al. 2011 ; Renahy et al. 2009 ; Vuitton et al. 2009), son objectif premier est de diminuer le temps (et donc le coût) nécessaire à la conception d'une LC. Automatiser ce processus implique deux types de contraintes : (1) liées au travail terminologique et (2) liées à la conception d'une LC (Renahy et al. 2009), à savoir recenser l'ensemble du lexique d'une LC (qu'il soit terminologique ou non), gérer les constructions particulières, notamment les structures lexicales, et respecter les principes communs à toute LC (non-ambigüité et non-redondance). Ceci présuppose la gestion de relations entre les unités lexicales (UL), qu'elles soient lexico-sémantiques (synonymie, antonymie), morphologique (flexionnelle, dérivationnelle, de variation morphosyntaxique faible, etc.) ou syntaxico-lexicales (grâce à la recherche de collocations par patterns prédictifs) (Plaisantin Alecu et al. 2012).

L'implémentation logicielle de ce processus, i.e. la Station Sensunique, automatise l'extraction d'UL candidates (ULC) à partir de corpus. Elle est configurable en finesse pour répondre aux multiples contextes d'utilisation possibles des ressources à construire, en termes de : domaines, types de textes, publics cibles des textes rédigés en une LC, ressources terminologiques préexistantes, ou plus généralement ressources linguistiques existantes et accessibles (notamment avec le courant fortement émergent des linked open linguistics data[2] (Chiarcos et al. 2012)). Elle offre aussi les fonctionnalités adéquates aux étapes suivantes du processus, à savoir, premières sélection et validation des UL par un analyste, seconde validation par l'expert métier et export de la ressource finale exploitable. Les interfaces utilisateurs (*interface de gestion* et *interface de travail*), très ergonomiques, ne nécessitent aucun savoir-faire technique et sont faciles à prendre en main et à explorer. L'application exploitant les lexiques d'une LC à concevoir (besoin initial de la Station) est un logiciel d'aide à la rédaction de textes techniques en LC sur mesure. La Station a été évaluée et validée dans ce cadre précis, sur l'intérêt du procédé de multi-extraction, implémenté dans la Station, pour le recensement du lexique d'une LC (Plaisantin Alecu et al. 2012).

Le processus métier d'acquisition du lexique d'une LC est très proche de l'acquisition de ressources termino-ontologiques (RTO) tel que décrit par Bourigault (2003). L'acquisition de dictionnaires, glossaires, lexiques, thesaurus[3] à partir de corpus doit répondre à une double contrainte de pertinence, vis-à-vis du corpus et de l'application visée, e.g. aide à la traduction, extraction d'information, indexation, etc. (Bourigault 2003).

La Station répondant à ces contraintes, son champ d'application initial (lexique d'une LC) peut être élargi à l'ensemble des RTO. Ses fondements méthodologiques et son architecture logicielle donne à la Station le potentiel d'un outil générique pouvant produire des ressources variées tout en étant fonction de l'application visée. Dans ce sens, elle suit le principe d'adéquation de Slodzian (2003) : *Qu'il s'agisse d'indexation, de mémoires de traduction bi- ou multilingues, d'aide à la rédaction de docu-*

2 Qu'il faudra toutefois télécharger puis convertir au format des ressources intégrables compatibles à la Station.

3 Pour la liste complète des RTO, voir Bourigault (2003).

ments experts, les outils proposés doivent présenter un degré d'adéquation suffisant avec le problème que l'utilisateur cherche à résoudre.

Dans la suite de cet article, nous allons tout d'abord situer notre travail (&1) dans le contexte des LC (&1.1) pour ensuite définir les besoins qui ont guidé la conception de la Station Sensunique (&1.2). Nous présenterons ensuite l'architecture générale de la Station Sensunique (&2), en mettant l'accent sur ses aspects modulaire et paramétrable. Nous décrirons les multiples possibilités de paramétrage de l'analyse automatique, en fonction du corpus d'entrée et de l'application visée. Puis nous présenterons le module de gestion qui, à partir de la liste des ULC issues du module d'analyse automatique, offre un ensemble de manipulations visant à faciliter les processus de sélection et de validation manuelle de ces ULC. Nous finirons par la présentation du module d'export qui permet également un paramétrage fin des ressources à constituer en fonction de l'application visée. Dans la conclusion, nous présenterons les possibles évolutions de la Station Sensunique.

Nous nommons « analyste » tout linguiste, terminologue, ingénieur des connaissances ou autre utilisateur de la Station et « ressources » tout type de RTO et lexique d'une LC.

1.1 Contexte

Les recherches concernant les Langues Contrôlées, sous leur multiples dénominations (langues simplifiées, langues construites, etc.) ne sont pas nouvelles, même si largement sur l'anglais : Kuhn (2013) recense plus de 100 LC conçues dans cette langue. Les travaux portant sur le français restent extrêmement rares : on peut citer l'initiative du COSLA (Comité de Simplification de la Langue Administrative ou le Français Rationalisé du GIFAS (Groupement d'Industries Françaises Aéronautiques et spatiales mais qui avait pour bases les règles du Simplified English de l'AECMA (Association Européenne des Constructeurs de Matériel aérospatial) (GIFAS, 1990).

Pour situer notre approche sur le panorama des travaux entrepris sur les LC, nous reprendrons une récente enquête sur l'ensemble des LC anglaises, dans laquelle Kuhn (2013 : 3) propose la définition suivante : *A controlled natural language is a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics while preserving most of its natural properties.*

Nous sommes en accord avec cette définition, dans le sens où elle met l'accent sur les caractéristiques essentielles des LC telles que nous les concevons : une LC est toujours construite à partir d'une langue naturelle et en conserve les propriétés. Si l'on adopte la classification des LC proposée par Kuhn (2013), les LC que nous concevons sont de type CTWDAI, puisque : elles sont conçues dans l'objectif d'améliorer la compréhensibilité (*Comprehensibility)* ; elles augmentent la traductibilité (*Translability*) ; elles sont destinées à être écrites (*Written*) ; elles sont spécifiques à un domaine (*Domain-dependent*) ; elles sont initiées par une recherche académique (*Academic*) ; elles sont aussi industrielles (*Industrial*) dans la mesure où l'applicabilité des LC en industrie est un critère prépondérant des travaux présentés dans cet article.

Plus qu'une LC, nous cherchons à mettre en place un cadre méthodologique de conception assistée de LC *sur mesure* . Nous appelons LC sur mesure une LC reposant sur les besoins précis d'une structure particulière ayant pour objectifs l'amélioration de la qualité de son système documentaire et l'amélioration de la fiabilité de ses textes afin de diminuer les risques liés à leur mauvaise interprétation/application. Une telle LC est donc circonscrite à un domaine et à un environnement de rédaction précis, c'est-à-dire une activité précise, un public défini et un type de textes particulier (Renahy et al. 2009). Elle repose sur une analyse de corpus délimité, lequel doit recenser l'ensemble des textes en vigueur pour l'activité et le public concernés[4] (Plaisantin Alecu et al. 2012). Enfin, la LC conçue doit permettre aux personnes en charge de la rédaction technique au sein d'une structure de rédiger des documents en conformité avec les principes de cette LC.

Le cadre méthodologique d'établissement d'une LC sur mesure doit intégrer une collaboration étroite entre les linguistes et les experts du domaine et de l'activité concernés (experts métier). Il doit, de plus, prendre en considération le coût et temps de conception. Ce temps a été estimé par Jeff Allen (2005) à 5 à 10 ans dans un contexte industriel. Jusqu'à aujourd'hui, ce coût de conception était un frein à l'exploitation de LC par des structures autres que les grandes industries[5]. L'accessibilité des LC à des structures plus modestes par la diminution de leur coût de conception a orienté nos travaux sur le cadre méthodologique de conception des LC.

Les travaux présentés ici concernent le premier verrou que nous avons souhaité lever, à savoir le recensement du lexique d'une LC sur mesure. La spécificité du recensement de ce lexique par rapport à la conception de ressources terminologiques est qu'il se doit d'être exhaustif : toutes les unités lexicales nécessaires lors de l'écriture effective de documents, qu'elles soient ou non terminologiques, doivent être encodées (pour être utilisables) dans le dictionnaire de la LC. Cependant, ce critère d'exhaustivité lexicale ne doit pas permettre n'importe quel emploi des unités recensées par le futur rédacteur : l'emploi de certaines unités lexicales doit être contraint. Ceci implique de distinguer au moins deux types de dictionnaires pour chaque LC sur mesure : un dictionnaire des unités lexicales d'une LC et un dictionnaire des structures lexicales, deux notions que nous précisons par la suite.

Le développement de la Station Sensunique s'inscrit dans cet objectif précis : accélérer l'acquisition du lexique pour la construction du Lexique d'une Langue Contrôlée (LLC).

1.2 Lexique d'une Langue Contrôlée: recensement des besoins

La notion du lexique d'une LC telle que nous la considérons mérite quelques précisions dans la mesure où elle ne correspond pas à la définition du 'vocabulaire contrôlé' ('lexique contrôlé', 'termes normalisés') communément employée dans la communauté scientifique :

4 Pour donner un exemple de corpus délimité, le système documentaire du laboratoire d'immunologie avec lequel nous avons travaillé comporte environ 40 modes opératoires.

5 Une liste récente des entreprises ayant investi dans la création d'une LC est donnée par Uwe Muegge et disponible ici : http://www.tekom.de/upload/2750/tcw02_2009.pdf [03/04/2014].

Vocabulaire contrôlé (ou lexique contrôlé ou liste de termes normalisés) : Un vocabulaire contrôlé est un ensemble de termes reconnus, fixés, inaltérables, normalisés et validés par un groupe (une communauté de pratiques) utilisés pour indexer ou analyser le contenu et pour rechercher de l'information dans un domaine d'information défini (...). [6]

La première différence entre un LLC et un vocabulaire contrôlé vient de son objectif. Comme le montre la définition précédente, un vocabulaire contrôlé est dans la majorité des cas défini pour l'indexation de documents dans le but d'en faciliter la recherche. Par exemple, le MeSH considéré comme vocabulaire contrôlé (Névéol, 2004) sert à l'indexation de ressources de santé.

La deuxième différence vient de leur périmètre respectif. L'ensemble des unités composant un vocabulaire contrôlé renvoie uniquement aux termes spécifiques d'un domaine. Le LLC, quant à lui, doit permettre la rédaction d'un texte technique dans sa globalité tout en respectant l'ensemble des contraintes d'une LC. Il devra donc recenser bien plus que les termes afin de pouvoir rédiger un texte en entier. En ce sens, (Møller et al. 2006) parle de « mots » (référant alors à des unités monolexémiques comme multilexémiques) afin de ne pas confondre les unités d'un LLC avec des unités terminologiques. Nous choisissons, quant à nous, de considérer comme *unités lexicales*[7] (UL) toutes les unités d'un LLC.

Pour recenser l'ensemble du vocabulaire contenu dans une collection de textes techniques, plusieurs types de vocabulaires sont nécessaires, comme le souligne également Camlong (1996). Ensemble, ils constituent un continuum allant du vocabulaire terminologique du domaine jusqu'au vocabulaire général. En effet, pour écrire un protocole dans le domaine d'immunobiologie[8], par exemple, plusieurs types de vocabulaire sont nécessaires :

- les termes du domaine (simples et complexes) : nominaux (*anticorps monoclonaux, réactif de lyse, tampon de fixation*), verbaux (*numéroter (les cellules), centrifuger (la suspension cellulaire)*) et adjectivaux (*aneuploïde, mononucléé*) ;
- les termes d'un autre domaine (*fenêtres informatiques, répartitions gaussiennes*) ;
- les unités du lexique général: soit entrant dans la composition des termes (*anticorps de souris*) ; soit 'autonomes' (*échantillons, divers, en particulier, étude*, etc.) ; soit potentiellement ambigües, puisque possédant un sens spécifique dans le domaine traité (*solution, population* dans le domaine de l'immunobiologie, par exemple).

Une LC exige le respect, entre autres, des principes de non-ambiguïté et de non-redondance : une unité lexicale ne peut avoir qu'une seule définition ; et une définition ne peut correspondre qu'à une seule unité lexicale dans un domaine choisi. Pour être en conformité avec ces exigences, il s'avère

6 http://appui.upmf-grenoble.fr/wiki/index.php/Vocabulaire_contrôlé [03/04/2014].
7 Nous reprenons ici la notion d'unité lexicale telle que définie par L'Homme (2005).
8 Tous nos exemples se basent sur le corpus-test établi dans le domaine d'immunobiologie, composé de 14 protocoles, pour un total de 10 064 mots. Ce corpus a été soumis à des extracteurs de termes, ce qui a permis de produire une liste de 2945 unités lexicales candidates (ULC), parmi lesquelles 1512 unités lexicales ont été finalement validées par un analyste.

nécessaire de contrôler l'ensemble du lexique utilisé pour la rédaction de la documentation dans un domaine. Pour exemple, il est indispensable d'éviter d'employer le mot *solution* au sens général (*Ensemble des opérations mentales, intellectuelles susceptibles de fournir une réponse théorique ou pratiques visant à la résolution, l'analyse, la compréhension d'un problème (...), TLFi*[9]) dans les protocoles d'immunobiologie, dans lesquels *solution* prend un sens très spécifique (*Liquide formé par la dissolution d'une substance solide (p. ex. médicament) dans un solvant*, GDT[10]). Il est également nécessaire d'identifier de multiples relations entre les unités lexicales ou leurs formes telles que :

- relations morphologiques (relation flexionnelle, relation dérivationnelle, relation de variation morphosyntaxique faible, etc.) ;
- relations lexico-syntaxiques (grâce à la recherche de collocations par patterns prédictifs) ;
- relations lexico-sémantiques (par exemple, relations de synonymie, homonymie).

1.2.1 Structures lexicales d'une Langue Contrôlée

Nous introduisons la notion de structure lexicale pour répondre au critère de non-ambigüité tout en conservant le caractère exhaustif du lexique et la nécessité de restriction d'emploi selon le contexte. La notion de structure lexicale dépasse la définition d'unité lexicale à strictement parler puisqu'elle s'appuie sur la combinatoire lexico-syntaxique entre plusieurs unités lexicales, se situant ainsi à la frontière du lexique et de la syntaxe. Cette notion est à rapprocher de celles de classes de sélection distributionnelles, classes d'objets, fonctions lexicales, cadres prédicatifs, pour ne citer que quelques unes des dénominations décrivant ces types de construction dans différentes théories linguistiques. On définit une Structure Lexicale (SL) comme un patron morphosyntaxique imposé et contrôlé par un lexème, souvent prédicatif, composée d'une partie figée (lexicalisée, variable uniquement en flexion) et d'une partie variable (mais contrainte par des traits morphosyntaxiques et sémantiques). Par exemple, *marquage* est le lexème prédicatif dans *marquage des cellules, marquage des cellules leucocytaires, marquage des cellules endothéliales vasculaires animales, marquage des cellules en suspension*. Le besoin de définir des structures lexicales vient, d'une part, de l'impossibilité d'encoder ces constructions dans un dictionnaire de termes (puisque ce ne sont pas des UL) et, d'autre part, de la nécessité de contrôler leur distribution et leur variabilité dans un environnement de rédaction d'une LC. C'est pour ces raisons que nous proposons de les recenser dans un dictionnaire spécifique, sous un format décrivant leurs principales caractéristiques :

Exemple

marquage de < *NOM : CELLULE*>

La partie variable, introduite par les chevrons (<>), est généralement définie par sa catégorie fonctionnelle (ici : NOM), qui peut être en plus caractérisée par son appartenance à une classe sémantique (ici : CELLULE).

9 Trésor de la Langue Française Informatisé, atilf.atilf.fr [03/04/2014].
10 Grand Dictionnaire Terminologique, http://gdt.oqlf.gouv.qc.ca/Resultat.aspx [03/04/2014].

La notion de structure lexicale est primordiale lorsque, nous éloignant de la théorie terminologique classique, nous considérons comme termes des syntagmes autres que les syntagmes nominaux. En effet, des verbes ou des adjectifs peuvent renvoyer à des concepts bien spécifiques dans des domaines précis. Certains dictionnaires terminologiques recensent d'ores et déjà des termes de nature verbale. Par exemple, on trouve aussi bien le nom 'centrifugation' que le verbe 'centrifuger' dans Le GDT. Simplement, la description de ce verbe, en s'arrêtant à l'identification de sa catégorie verbale, ne nous renseigne ni sur la présence ni sur la nature de ses compléments : pourtant, on *centrifuge* toujours *quelque chose, du sang total, du plasma sanguin* etc. Nous proposons donc de recenser ce verbe dans un dictionnaire de structures, en indiquant clairement qu'il doit être accompagné de compléments d'une certaine classe fonctionnelle et sémantique : centrifuger <NOM : SANG> .
Un autre avantage concernant l'identification des structures lexicales est l'établissement des relations entre des UL dérivées et la vérification de la cohérence du recensement du vocabulaire. En théorie, les UL prédicatives en relation de dérivation ne peuvent introduire dans leurs structures que des compléments appartenant à des classes sémantiques identiques :

Exemple
numéroter < NOM : CELLULE> ; <NOM : CELLULE> numéroté(es) ; numération de <NOM : CELLULE>
Pour rédiger *: numération des populations leucocytaires, numéroter les lymphocytes T, B et NK*

L'avantage du recensement de ces structures est double : d'une part, cela permet de contrôler que *populations leucocytaires* et *lymphocytes T, B et NK* portent bien la contrainte sémantique *CELLULE* et que *numéroter, numération* (voire le participe passé adjectival *numéroté*) renvoient toujours à la même classe sémantique.
En résumé, le recensement du LLC implique la création de dictionnaires pour quatre types de données : les unités lexicales terminologiques, les unités lexicales non-terminologiques, les structures lexicales terminologiques et les structures lexicales non-terminologiques.

2 Architecture générale de la Station Sensunique

A notre connaissance, il n'existe pas d'outil spécifique dédié à l'aide au recensement du LLC. Par contre, il existe de nombreux outils d'extraction de termes, tâche à laquelle s'apparente l'établissement du LLC. Par conséquent, après avoir testé l'hypothèse que les extracteurs de termes peuvent aider au recensement d'un LLC (Plaisantin Alecu et al. 2012), nous avons décidé de les intégrer à la Station Sensunique.

Comme toute plateforme terminologique (par exemple : HyperTerm[11], Terminae[12], Terminus[13]), la Station intègre la mise en séquencement de plusieurs outils TAL (étiquetage, lemmatisation et extraction de termes). Sa spécificité repose sur ses autres fondements méthodologiques. Le premier est la multi-extraction ou coopération de plusieurs extracteurs. Ce procédé donne des résultats significativement meilleurs que l'utilisation d'un seul extracteur et il permet de réduire le silence et filtrer automatiquement le bruit. Plus précisément, cumuler les résultats de 3 extracteurs de termes permet de couvrir 79 % des termes (par opposition à 58% de rappel pour le meilleur extracteur), et le meilleur moyen d'aider à déterminer le statut terminologique d'une ULC est de se baser sur les résultats communs aux 2 extracteurs (Yatea et Termostat dans l'étude) avec une précision de 37 % par opposition à 28% d'un seul extracteur (Plaisantin Alecu et al. 2012). Ce procédé reprend celui des systèmes à base de vote (Fiscus 1997 ; Brunet-Manquat 2004 ; Matusov 2007 ; Serp et al. 2008), mais n'a jamais été employé avant nos travaux pour l'acquisition de ressources.

La seconde spécificité de la Station est le recoupement des résultats d'extraction avec des ressources lexicales et terminologiques existantes interrogées automatiquement. Ceci permet, d'une part, d'augmenter le potentiel terminologique d'une ULC déjà recensée comme terme dans une ressource externe, et d'autre part d'attribuer un statut non-terminologique à des ULC présentes dans les ressources lexicales intégrées à la Station.

Le dernier fondement méthodologique est le calcul de trois pondérations, en fonction de diverses informations recueillies automatiquement par la Station : (1) le Poids Terminologique (PT) ou potentiel d'une ULC à être un terme ; (2) le Poids de Structure Lexicale (PSL) ou potentiel d'une ULC à être transformée en une structure lexicale ; et (3) le Poids d'Unité Lexicale (PUL) ou potentiel d'une ULC à être une unité lexicale bien formée. Le calcul de ces pondérations organisent le travail de validation et facilitent la prise de décision et l'établissement de consensus entre plusieurs analystes ou entre l'analyste et l'expert métier.

Bien que chacun de ces procédés (multi-extraction, interrogation des ressources existantes, pondération) ne soient pas nouveaux, ils n'ont jamais été combinés, à notre connaissance, pour cumuler leurs bénéfices au sein d'une seule et même plateforme de recensement de ressources terminologiques ou non terminologiques.

La Station s'articule sur deux points de vue du processus d'acquisition de ressources : (1) chronologique (centré processus) : import des textes d'entrées, analyse automatique[14], validation, et enfin export ; (2) ergonomique (centré analyste) : mise en adéquation de l'analyse selon le corpus et l'application visée par la ressource, visualisation des ULC (fiche lexicale et contextes d'occurrence), analyse d'un groupe d'ULC (pour l'organisation du travail ou pour des actions en masse pertinentes), re-

11 http://www.tedopres.com/hyperterm-terminology-management [03/04/2014].
12 http://lipn.univ-paris13.fr/terminae/index.php/Main_Page [03/04/2014].
13 http://terminus.iula.upf.edu//cgi-bin/terminus2.0/terminus.pl [03/04/2014].
14 L'analyse automatique comprend : étiquetage, lemmatisation, racinisation, extraction des ULC, interrogation des ressources externes et internes, calcul des pondérations.

cherches complexes en corpus ou dans la liste d'ULC, modification ou enrichissement des descriptions ou relations des ULC, validation progressive, demande de validation par l'expert-métier, etc.
De cette double articulation résultent 4 modules distincts : (1) Module de configuration de l'analyse automatique, (2) Module d'analyse automatique, (3) Module de gestion des ULC et (4) Module d'export, ainsi que 3 étapes successives d'établissement du lexique (Figure 1) :

- Etape 1 : Analyse automatique, qui extrait, à partir d'un corpus textuel, une liste composée d'unités terminologiques et non-terminologiques classées en fonction de leur statut et de leur potentiel terminologique ;
- Etape 2 : Analyse manuelle approfondie, qui consiste en un premier filtrage de la liste opéré par l'analyste pour ne retenir que les unités potentiellement valables et un second filtrage réalisé avec l'aide de l'expert métier aboutissant à des ressources validées ;
- Etape 3 : Etablissement et export paramétré des ressources établies.

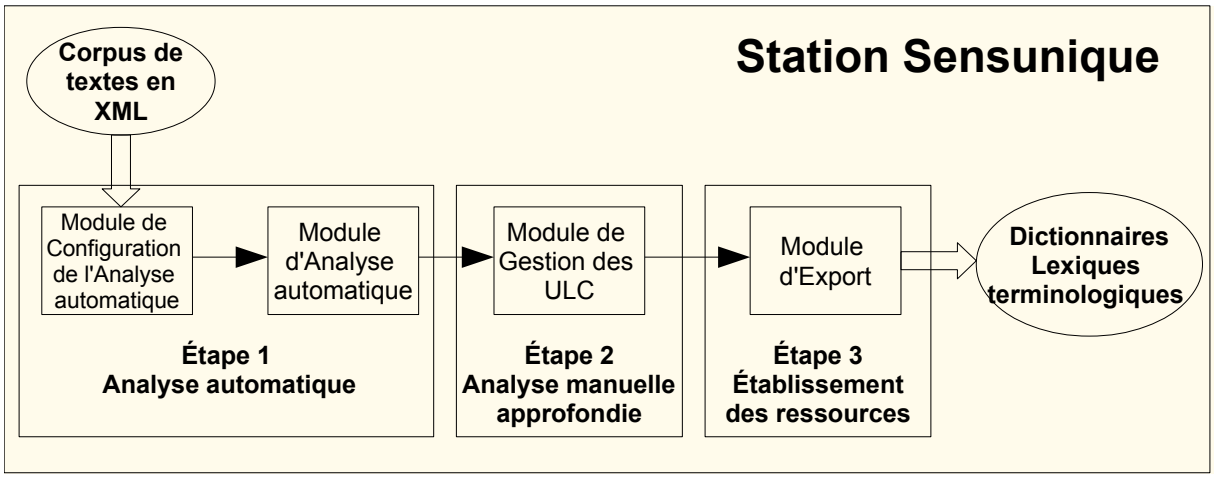

Figure 1: Schéma de la Station Sensunique.

2.1 Module Configuration de l'analyse automatique: Paramétrer l'analyse en fonction de la ressource visée

L'analyse automatique doit être configurée en fonction de l'application visée. L'analyste peut choisir ce qu'il souhaite exploiter comme types de corpus, outils, ressources et valeurs initiales de l'algorithme de pondération, selon leur adéquation au corpus et à la ressource visée. La qualité des résultats de l'analyse-extraction dépend de ces paramètres.

2.1.1 Sélection de corpus

Pour le même projet, trois types de corpus textuels[15] peuvent être simultanément analysés par la Station :

(1) le Corpus d'Analyse (CA) : c'est un corpus obligatoire duquel sont extraites les ULC à analyser ;

(2) le Corpus Support (CS) : c'est un corpus facultatif, du même domaine que le CA. En recoupant les ULC retrouvées dans les deux corpus (CA et CS), l'algorithme de pondération renforce leur potentiel terminologique. Ce procédé est inspiré de l'hypothèse de Drouin (2003) prouvant qu'une UL extraite de deux corpus différents du même domaine a plus de probabilité d'être un terme du domaine ;

(3) le Corpus Contrastif (CC) : c'est un corpus facultatif, contenant des textes généralistes, non relatifs au domaine analysé. L'exploitation d'un CC permet à l'algorithme de pondération d'augmenter la qualité des résultats en diminuant le potentiel terminologique des ULC issues du CA et du CC à la fois. De nouveau, ce procédé est inspiré de Drouin (2003) qui prouve qu'une UL extraite d'un corpus de domaine et d'un corpus généraliste a plus de probabilité d'être une unité du lexique général qu'un terme du domaine.

Les corpus sont (ré)utilisables dans plusieurs projets. En outre, un corpus n'est pas intrinsèquement lié à un statut particulier (CA, CS ou CC) : ce statut lui est attribué en fonction du projet, par un analyste. Par conséquent, le même corpus peut être utilisé comme un CA dans un projet particulier et comme un CC dans un autre projet. Ceci permet une meilleure exploitation de différents corpus constitués dans un groupe de travail ayant des projets différents.

2.1.2 Sélection des outils

Pour effectuer une analyse automatique, la Station intègre un certain nombre d'outils, à savoir :

- les étiqueteurs morphosyntaxiques : statistique Treetagger (Schmid, 1994) et à base de règles Brill[16] (Brill 1992) ; l'annotation de chaque forme fléchie du corpus par sa catégorie morphosyntaxique et ses traits morphosyntaxiques est utile non seulement à l'analyse flexionnelle et à l'extraction de termes mais également aux diverses recherches en corpus que l'analyste peut effectuer via le concordancier intégré à la Station Sensunique ;
- l'analyseur flexionnel du français Flemm v2 et v3 (Namer, 2000) : l'annotation de chaque forme fléchie du corpus par sa formé lemmatisée est utile non seulement aux extracteurs mais également aux diverses recherches en corpus que l'analyste peut effectuer via le concordancier intégré à la Station Sensunique ;
- les extracteurs de termes Acabit (Daille, 1994), TermoStat (Drouin, 2003) et YaTeA (Aubin et al., 2006) : les extracteurs de termes fournissent chacun des propositions de termes assortis d'une ma-

15 Mis au préalable au format XML TEI P5, http://www.tei-c.org/Guidelines/Customization/Lite/teiu5_fr.html [04/04/2014].

16 Avec le lexique et le fichier de règles fournis par l'ATILF-CNRS, de Nancy.

trice morphosyntaxique ; de plus, Acabit regroupe des variantes du même terme ; Acabit et YaTeA découpent les termes composés en tête et expansion ; enfin, d'autres informations fournies par les extracteurs permettent de calculer certaines types de collocations (ULC incluses, composées et associées, cf. & 2.2) ;

- le racinisateur Lingua::Stem[17]: les racines ajoutées grâce à cet outil permettent d'identifier les relations dérivationnelles entre les ULC et sont également exploitées pour une recherche en corpus via le concordancier.

Les outils sont reliés en chaînes de travail indépendantes et parallèles. L'analyste peut sélectionner de 1 à 3 chaînes d'outils parmi : (1) TreeTagger - Termostat ; (2) Brill - Flemm v2 - Acabit ; (3) TreeTagger - Flemm v3 - YaTeA. Bien que la sélection d'une seule chaîne suffise pour lancer une analyse automatique, la Station est optimisée lors de l'emploi des 3 chaînes grâce au procédé de multi-extraction. Les résultats d'analyse de toutes les chaînes sélectionnées sont cumulés et recoupés et les informations obtenues affichées dans la liste des ULC résultant de l'analyse.

2.1.3 Sélection de ressources terminologiques externes (prédéfinies)

Deux ressources externes sont actuellement prédéfinies dans la Station :

- TermSciences[18], portail terminologique multidisciplinaire développé par CNRS-INIST ;
- IATE[19], base de données terminologique de l'Union Européenne.

L'interrogation automatique par web service de ces deux ressources externes permet de vérifier si une ULC proposée par les extracteurs est déjà recensée en tant que terme. Pour IATE, l'interrogation peut être restreinte à un domaine ou un sous-domaine précis (selon le référencement en domaines et sous-domaines EuroVoc[20]). Seuls les termes qui atteignent une certaine fiabilité (selon le paramètre «reliability» défini par IATE) sont retenus. Pour TermSciences, l'interrogation permet de vérifier si les constituants d'une ULC composée (sa tête ou son expansion) sont recensés indépendamment comme terme.
L'interrogation des ressources externes influe sur les pondérations, en renforçant le potentiel terminologique d'une ULC attestée dans une (ou plusieurs) ressource(s), renforcement plus ou moins fort selon si l'ULC est attestée dans sa globalité, ou si sa tête et /ou son expansion sont attestés. Elle permet ainsi de structurer le processus de validation des ULC. De plus, elle participe à l'enrichissement des informations rattachées à chaque ULC, puisque sont importées dans la Station des informations supplémentaires telles que définitions, synonymes et classes sémantiques/conceptuelles auxquelles appartient le terme attesté.
L'analyste peut choisir d'intégrer ou non l'interrogation automatique des ressources à l'analyse.

17 http://search.cpan.org/~sdp/Lingua-Stem-Fr0.02/lib/Lingua/Stem/Fr.pm [04/12/2011].
18 http://www.termsciences.fr/ [03/04/2014].
19 http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load [03/04/2011].
20 eurovoc.europa.eu [03/04/2011].

2.1.4 Intégration de nouvelles ressources (dites internes)

En plus de ressources externes prédéfinies, la Station permet d'intégrer à chaque nouveau projet d'autres ressources spécifiques, moyennant leur mise au format prédéfini dans la Station. Il peut s'agir aussi bien de ressources terminologiques (e.g. des dictionnaires spécialisés) qui augmentent le potentiel terminologique des ULC, que des ressources non-terminologiques (e.g. Morphalou 2.0[21]) qui augmentent le poids d'unité lexicale d'une ULC tout en diminuant son potentiel terminologique. Par ailleurs, des ressources constituées au préalable dans la Station, résultant d'autres projets, peuvent aussi être intégrées en tant que ressources internes.

Du fait de l'intégration dynamique des ressources, la Station peut être considérée comme évolutive, puisque chaque analyse peut être enrichie grâce à un ensemble de ressources spécifiques et appropriées.

2.1.5 Paramétrage des pondérations

Trois pondérations servent à faire ressortir la fiabilité des ULC et à les classer en vue d'organiser le travail de filtrage et de validation :

(1) Poids Terminologique (PT) : potentiel terminologique d'une ULC calculé selon 7 critères :

- le nombre des extracteurs ayant proposé l'ULC ;
- le seuil du statut terminologique, c'est-à-dire la valeur à partir de laquelle les ULC sont considérées comme termes ;
- présence dans le CS ou le CC (cf &2.1.1) ;
- le nombre des ressources choisies ayant attesté l'ULC ;
- le type d'attestation dans une ressource (l'attestation d'ULC globale ayant plus de poids que l'attestation de la tête et/ou l'expansion seulement) ;
- la fiabilité de la ressource externe (TermSciences ou IATE)[22] dans le domaine analysé ;
- et la présence d'une ULC dans une ressource terminologique interne (cf. & 2.1.4).

(2) Poids de Structure Lexicale (PSL) : potentiel d'une ULC à être transformée en une structure lexicale, calculé selon 8 critères dont :

- l'attestation d'une ULC globale dans une ressource terminologique (qui influe négativement sur sa possibilité d'être une structure lexicale) ;
- la matrice morphosyntaxique d'une ULC (les verbes et les participes ayant plus de probabilité de constituer les structures lexicales) ;
- le nombre de dérivées et/ou de collocations construites autour d'une ULC.

21 Lexique de formes fléchies du français développé par ATILF, http://www.cnrtl.fr/lexiques/morphalou/LMF-Morphalou.php [03/04/2014].

22 Estimée par l'analyste.

(3) Poids d'Unité Lexicale (PUL) : potentiel d'une ULC à être une unité lexicale bien formée, calculé selon 2 critères :

- le nombre d'extracteurs l'ayant proposé ;
- la présence d'une ULC dans une ressource interne non-terminologique (cf. & 2.1.4).

A chacun de ces critères correspond une valeur jouant dans le calcul global de chacune des 3 pondérations. Des valeurs préexistent par défaut, mais sont ajustables par l'utilisateur.

2.2 Module d'Analyse automatique

Ce module a deux fonctions. La première fonction est d'annoter linguistiquement le corpus d'analyse par incorporation des résultats des étiqueteurs, lemmatiseurs et racinisateur intégrés. Sa deuxième fonction est d'extraire de ce corpus des ULC (par multi-extraction), de les décrire (résultat des extracteurs et de l'interrogation des ressources définies) et de les pondérer (résultat de l'algorithme de pondération de la Station).
Les informations issues de l'analyse automatique sont, pour chaque ULC :

- Forme canonique : correspond, la plupart de temps, à la suite de lemmes de chaque élément d'une ULC, ex. *membrane cellulaire* ;
- Statut lexical : terminologique ou non, selon le seuil du PT paramétré par l'analyste ;
  > Domaine(s) (uniquement si le statut est terminologique ; correspond dans ce cas au domaine renseigné par l'analyste dans le descriptif du projet ; ex. *immunobiologie*) ;
- Usage : "préconisé" ou "interdit", selon les spécifications d'une LC ;
- Catégorie(s) sémantique(s) : proposée(s) par les ressources externes (ex. *Structures cellulaires,* d'après TermSciences) ;
- Fréquence : nombre d'occurrences des formes fléchies de l'ULC en corpus ;
- Indices de confiance :
  > Pondérations internes : PT, PSL, PUL (cf & 2.1.5) ;
  > Indices des extracteurs externes : indices de confiance fournis par les extracteurs, ex. *loglike* pour Acabit ;
- Tête : régisseur syntaxique d'une ULC, ex. *membrane* ;
- Expansion : complément/modifieur d'une Tête, ex. *cellulaire* ;
- Catégorie morphosyntaxique fonctionnelle: en général, catégorie de la Tête d'une ULC, ex. *NOM* ;
- Matrice morphosyntaxique : suite des catégories morphosyntaxiques de chaque élément de l'ULC., ex. *Adj Nom* ;
- Formes fléchies : si trouvées en corpus, assorties des traits morphosyntaxiques et fréquence ;
- Variantes : provenant soit du corpus analysé, soit des ressources externes, ex. *membrane plasmique* ;
- ULC dérivées: ULC dont un des composants appartient à la même famille dérivationnelle, ex. *membrane cellulaire, marquage de cellule* ;

- ULC homonymes: ULC homographes d'une autre catégorie morphosyntaxique que l'ULC analysée;
- Collocations (ULC liées) ;
  > ULC incluses : une ULC incluse est une ULC dont l'intégralité se retrouve dans l'ULC analysée ; par exemple, pour l'ULC *anticorps monoclonal de souris*, les ULC incluses sont : *anticorps monoclonal, anticorps* ;
  > ULC composées : une UL composée est une ULC contenant plus que l'intégralité de la ULC analysée ; par exemple pour l'ULC *anticorps monoclonal*, les ULC composées sont *anticorps monoclonal conjugué, anticorps monoclonal de souris, anticorps monoclonal HLA-B27* [23];
  > ULC associées : une ULC associée est une ULC non incluse et non composée contenant un même lemme que l'ULC analysée ; exemple : pour l'ULC *anticorps monoclonaux*, ULC associée est *solution d'anticorps* ;
- Sources :
  > Outil(s) ayant proposée une ULC (exemple : Termostat, Acabit) ;
  > Ressource(s) externe(s) l'attestant (exemple : TermSciences) ;
- Définition(s) (provenant de ressources externes).

Partant du principe que chaque proposition faite lors d'une analyse automatique peut être modifiée, tous les résultats du module d'analyse (excepté les indices de confiance calculés par les extracteurs et les sources) sont éditables dans le module de Gestion des ULC.

2.3 Module de Gestion des ULC: Faciliter le processus de sélection et de validation

Le module de gestion des ULC rassemble des fonctionnalités facilitant la seconde phase du processus d'acquisition des ressources, à savoir l'analyse manuelle approfondie. Elle consiste en un premier filtrage des ULC par un analyste et en l'établissement du consensus final avec les experts métier (Fig.1). Le parti pris fondamental de la Station est que l'analyste peut effectuer tout changement nécessaire concernant l'ensemble de résultats proposés par l'analyse automatique. Un espace dédié, appelé *interface de travail* lui sert à visualiser, à approfondir et à élargir (si besoin) les résultats afin de les valider pour construire la ressource finale.

Dans l'interface de travail, les résultats de l'analyse automatique peuvent être visualisés sous 3 modes :

- liste des ULC contenant des informations utiles pour trier et filtrer les résultats ;
- fiche lexicale de chaque ULC détaillant toutes les informations ;

23 Les UL incluses et composées fonctionnent de manière symétrique : si une ULC1 est ULC incluse d'une ULC2, alors l'ULC2 sera ULC composée de l'ULC1.

- fiches de relations, détaillant l'ensemble de relations entre l'ULC analysée et d'autres ULC (telles que variantes, collocations, homonymes, ULC appartenant à la même famille dérivationnelle).

L'analyste peut ajouter, modifier, compléter, valider ou supprimer toute ULC ou information à partir d'un mode de visualisation approprié. Chaque proposition/modification de données est toujours tracée, c'est-à-dire, assortie du nom de son auteur (qu'il soit analyste, outil ou ressource).
Ce module réunit également des fonctionnalités d'exploration (des ULC et de leurs informations descriptives) et d'aide à la décision (aux rejet, modification, enrichissement, validation):

- **tri et filtre** sur la liste des ULC selon 21 paramètres différents, dont fréquence, PT, PUL, extracteur(s) d'origine, ressources attestant l'ULC, matrice morphosyntaxique, catégorie sémantique etc. ; les filtres sont cumulatifs, c'est-à-dire qu'on peut filtrer les ULC selon plusieurs paramètres à la fois (par exemple, ULC proposées par Termostat, ayant atteint un certain seuil de PT et d'une matrice morphosyntaxique particulière) ;
- **projection** pour visualiser une ou plusieurs ULC en contexte d'origine (en corpus ou par phrases) ;
- **regroupement** d'ULC dans les fiches de relations ; certaines ULC sont regroupées automatiquement, mais l'analyste peut aussi établir de nouvelles relations ;
- **concordancier évolué** offrant différents types de recherche sur le corpus[24] : (a) simple : sur une chaîne de caractères ; (b) morphologique simple : sur un (ou une suite de) lemme(s) permettant d'identifier toutes ses formes fléchies d'une ULC ; (c) morphologique complexe : sur un (ou une suite de) radical(aux) permettant d'identifier les familles dérivationnelles ; (d) morphosyntaxique : sur une suite d'étiquettes morphosyntaxiques ; (e) recherche dite combinée permettant de coupler les types de recherches précédents. Combiner des critères appartenant à différents niveaux d'analyse linguistique 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 '[e]Nom [c] de [l] cellule' (exprimée sous forme d'expression régulière Sensunique) permet de cibler les groupes dont le premier élément est le Nom suivi de la préposition 'de' et d'une forme fléchie du mot 'cellule' (ex. *nombre de cellules, greffon de cellules, analyse de cellules* etc.).

L'établissement des SL se fait manuellement, à partir du regroupement de plusieurs ULC. La fonctionnalité de dégradation permet de définir une nouvelle SL (et ses différentes informations associées, telles que statut lexical, catégorie sémantique, catégorie fonctionnelle etc.) et de l'ajouter à une liste des SL. Les opérations de tri et de filtrage peuvent être effectuées sur la liste des SL comme sur la liste des ULC.
Enfin, 7 statuts de validation, correspondant à différentes étapes d'analyse ('Non validé', 'En cours d'analyse', 'A valider par les experts', 'Invalidée par les experts', 'Validé par les experts', 'Validée', 'Invalidé') permettent de suivre le processus d'établissement du lexique.

24 Sous forme d'Expressions Régulières (selon http://fr2.php.net/manual/fr/book.pcre.php) adaptées à la Station Sensunique.

2.4 Module d'Export: Paramétrer les ressources produites en fonction d'une application

Ce module permet d'exporter en dictionnaires les données recensées dans la station au format XML afin de :

- créer des ressources terminologiques diverses ;
- exploiter les données dans d'autres applications ;
- durant l'analyse, valider les données nécessitant des compétences spécifiques par des experts métiers.

En fonction de son objectif, l'utilisateur peut paramétrer les dictionnaires de sortie, en choisissant le(s) type(s) d'informations qu'il souhaite exporter. Toute la finesse de description d'une ressource produite dans la Station n'est pas forcément utile à l'application qui va exploiter cette ressource. De même, on peut n'être intéressé que par un périmètre restreint des UL recensées.
La sélection s'effectue à l'aide des filtres cumulatifs servant à restreindre le périmètre des données exportées selon deux axes :

- sélection des propriétés des ULC (parmi les 17 propriétés proposées, telles que définition, synonymes, matrice morphosyntaxique, catégorie sémantique, colocations, statut de validation, etc.) :

Exemple : dictionnaire d'UL contenant seulement : Forme canonique, Définition et Variantes
Exemple : dictionnaire d'UL contenant seulement : Forme canonique, Matrice morphosyntaxique et Fréquence

- sélection des propriétés des ULC et des valeurs de propriétés :

Exemple : dictionnaire d'UL contenant seulement : Forme canonique, Classe Sémantique, Définition, Statut de Validation ; ET le Statut de Validation est « Validée »

Le même projet permet de créer plusieurs ressources en fonction d'une application visée. Le principe est le même pour les dictionnaires de SL.

3 Conclusion

Conçue dans l'objectif d'optimiser (en termes de qualité et de coût) l'acquisition du lexique d'une langue contrôlée, les possibilités d'exploitation de la Station Sensunique dépassent considérablement ce champ d'action. En effet, l'éventail des configurations d'analyse (choix des outils et ressources, intégration de nouvelles ressources, personnalisation des bases de calcul des pondérations, paramétrage de l'export) en fonction de nombreux contextes d'utilisation, fait d'elle non seulement un outil d'acquisition du lexique d'une langue contrôlée, mais aussi une plateforme pertinente pour tout travail de constitution de RTO à partir de corpus.

Sur le plan méthodologique, la multi-extraction permet à la Station Sensunique d'offrir à ses utilisateurs les points forts de chaque extracteur de termes. Renforcée par l'interrogation des ressources existantes et par le principe des 3 types de corpus, la Station pondère ses résultats et permet ainsi d'organiser le processus de validation des ULC. L'interrogation des ressources existantes permet d'enrichir automatiquement la description morphologique, syntaxique et sémantique des ULC. Par ailleurs, la Station est conçue pour respecter et faciliter le processus métier d'acquisition de ressources : elle prend en compte les différentes phases de ce processus et modélise l'implication de plusieurs acteurs, y compris la validation finale par un expert-métier. De plus, 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. Enfin, la Station Sensunique est dotée d'une interface utilisateur facile à manier et à explorer.

En ce qui concerne les futurs développements de la Station, plusieurs directions sont envisagées. Premièrement, nous considérons l'ajout d'autres chaînes d'outils ou le développement d'outils propres pour améliorer les performances de la Station, ainsi que l'intégration d'autres ressources externes, bien que ceci pose problème concernant les licences d'utilisation des fois difficiles à obtenir : d'où l'importance d'interagir avec les courants tels que linked open data. Deuxièmement, nous souhaitons améliorer le traitement du contenu sémantique des textes, principalement la détection des relations sémantiques et conceptuelles entre les unités lexicales. Une autre direction de recherche est l'exploitation de la plateforme pour la construction d'ontologies de domaine.

4 Références bibliographiques

Allen J. (2005). How are we responding to industrial and business needs for Controlled Language and Machine Translation, *Journées Linguistiques – Langues contrôlées, traduction automatique et langues spécialisées : 5-6 May 2004 Besançon, France*, http://web.science.mq.edu.au/~rolfs/controlled-natural-languages/papers/Jeff-Allen.pdf [08/04/2014].

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, 2006, 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.

Camlong (1996). Méthode d'analyse lexicale textuelle et discursive, Paris, Orphrys.

Bourigault, D., & Aussenac-Gilles, N. (2003). Construction d'ontologies à partir de textes. In *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues* (pp. 27-50).

Brunet-Manquat F. (2004). Fusionner pour mieux analyser : Conception et évaluation de la plate-forme de combinaison. In *Actes de TALN-2004*. Fez, Maroc, 19-22 avril 2004. vol. 1/1, pp. 111-120.

Chiarcos Ch., Hellmann S. and Nordhoff S. (2012). Linking linguistic resources: Examples from the Open Linguistics Working Group, In Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), *Linked Data in Linguistics. Representing Language Data and Metadata*, Springer, Heidelberg, p. 201-216.

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.

Fiscus J.G.(1997). A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER), Automatic Speech Recognition and Understanding, 1997. In *Proceedings IEEE Workshop*, pp.347-354.

GIFAS, Guide du rédacteur. Groupement des Industries Françaises Aéronautiques et Spatiales, Paris, France, 1990.

Kuhn T. (2013) A Principled Approach to Grammars for Controlled Natural Languages and Predictive Editors. *Journal of Logic, Language and Information*, 22(1).

Kuhn T. (2014) A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40(1), 2014.

L'Homme M-C. (2005). Sur la notion de terme. In Meta: journal des traducteurs / Meta: Translators' Journal, vol. 50, n° 4, p. 1112-1132. http://id.erudit.org/iderudit/012064.

Matusov E. et al. (2007). System combination for machine translation of spoken and written language. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1222–237.

Møller M.H., Christoffersen E., Hansen M. (2006). Building a Controlled Language Lexicon for Danish. In *LSP and Professional Communication*, vol. 6, Nr. 1, p. 12-38.

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

Névéol A. (2004). Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé. In *RECITAL 2004*, Fès.

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., Betbeder M.L., Renahy J., Vuitton DA.(2012). *Optimisation d'un logiciel pour la rédaction de textes techniques de qualité : application-pilote au domaine de la santé*. Projet ANR -EMMA-2010-039 (2010-12), rapport final (non-publié).

Serp C., Cazal E., Laurent A., Roche M. (2008).TERVOTIQ : un système de vote pour l'extraction de la terminologie d'un corpus en français médiéval. In *9èmes journées internationales d'analyse statistique de données textuelles (JADT'2008),* Lyon, 2008.

Slodzian, M. (2000). L'émergence d'une terminologie textuelle et le retour du sens. Le sens en terminologie, 61-85.

Vuitton DA., Aishan A., Renahy J., Jin G., Wu X., De Grivel V., Cardey S. (2009). Controlled language: a Linguistic Concept to Improve Health Care Safety in a "Globalised" World? Application to Medical Proto-

cols Written within the Hospital Accreditation/Certification Framework in France and China. In *ISMTCL Proceedings, International Review BULAG*, PUFC, ISBN 978-2-84867-261-8, pp. 260-268.

Remerciements

Nos travaux ont été financés par l'Agence Nationale de la Recherche, programme Emergence 2010. Nous remercions toute l'équipe du projet Sensunique, les auteurs des outils intégrés et les organismes gérant les ressources terminologiques prédéfinies dans la Station Sensunique.