ÉCRIRE EN LANGUES SPECIALISEES : METHODES ET OUTILS DU TRAITEMENT AUTOMATIQUE DES LANGUES AU SERVICE DE L'AUTONOMIE DES REDACTEURS

I. Thomas

UNIVERSITE DE FRANCHE-COMTE

ECOLE DOCTORALE « LANGAGES, ESPACES, TEMPS, SOCIETES »

Mémoire en vue de l'obtention de l'Habilitation à Diriger les Recherches

ÉCRIRE EN LANGUES SPECIALISEES : METHODES ET OUTILS DU TRAITEMENT AUTOMATIQUE DES LANGUES AU SERVICE DE L'AUTONOMIE DES REDACTEURS

Présentée et soutenue publiquement par

Izabella THOMAS

Le 7 décembre 2016, à Besançon

Membres du jury :

Krzysztof BOGACKI, professeur à l'université de Varsovie, Pologne (président du jury)

Pierrette BOUILLON, professeur à l'Université de Genève, Suisse

Sylviane CARDEY- GREENFIELD, professeur à l'Université de Franche-Comté (rapporteur)

Natalie KÜBLER, professeur à l'Université Paris Diderot (rapporteur)

Christophe ROCHE, professeur à l'Université de Savoie (rapporteur)

Dominique Angèle VUITTON, professeur (émérite) à l'Université de Franche-Comté

Avant de commencer…

Le ‘nous’ que j’utilise dans ce mémoire n’est nullement un ‘nous’ de modestie. Cela fait plusieurs années que je ne travaille plus seule sur les projets de recherche. La recherche pour moi est devenue une activité collaborative réunissant plusieurs personnes qui mettent en commun leurs compétences et leurs intérêts. C’est de ce partage que la recherche tire sa force et il serait inconcevable pour moi de m’en approprier les résultats. Ainsi, derrière chaque ‘nous’ se cache au moins un étudiant en Master, un collègue, un collaborateur ou une équipe de projet. C’est eux que je voudrais remercier de m’avoir permis d’arriver là où je suis aujourd’hui. Je me permets de leur dire toute la satisfaction que je retire de ces collaborations et de leur rendre hommage en associant leur nom à ce mémoire.

Merci à Iana Atanassova, Mohand Beddar, Marie-Laure Betbeder, Sabeha Bichlé, Oleg Blagoskonov, Sylviane Cardey, Dilber DeVitre, Aleksandra Dziadkiewicz, Anastasia Galmiche, Bérenger Germain, Peter Greenfield, Gan Jin, Kyoko Kuroda, Lucie Laroche, Dalila Limame, Ziad Mikati, Thao Phan Thi Thanh, Blandine Plasaintin-Alecu, Julie Renahy, François- Claude Rey, Dominique Rota-Betain, Estelle Seillès, Igor Skouratov, Morgane Tornaboni, Dominique Angèle Vuitton, Xiaohong Wu, tous les chercheurs du Centre Tesnière et tous les étudiants dont j’ai eu le plaisir de diriger les mémoires.

Je remercie par avance les membres du jury pour le temps et l’attention qu’ils vont consacrer à ce mémoire, ainsi que pour les conseils et les perspectives qu’ils me donneront.

Un grand merci à mes ami(e)s-correcteurs-chercheurs Julie Renahy et Jovan Kostov.

Et aussi à Jean-Charles, qui, j’en suis certaine, est devenu le plus linguiste de tous les metteurs en scène de théâtre en France.

TABLE DES MATIERES

I. LISTE DES FIGURES 5

II. LISTE DES TABLEAUX 6

III. LISTE DES ABREVIATIONS 7

1. INTRODUCTION 8

1.1 TEXTES SPECIALISES ET CONTEXTES DE REDACTION 8
1.2 LANGUES SPECIALISEES ET LINGUISTIQUE 9
1.3 LEXIQUES SPECIALISES ET CONTEXTUALISATION 11
1.4 VERS LES BESOINS DES UTILISATEURS 12
1.5 OUTILLER LA REDACTION 13
1.6 VERS L'AUTONOMIE DES UTILISATEURS 15
1.7 ORGANISATION DU MEMOIRE 16

2. CONTEXTES DE REDACTION EN LANGUES SPECIALISEES 17

2.1 ÉCRITS PROFESSIONNELS (REDACTION TECHNIQUE) 18
2.1.1 PROJET LISE 18
2.1.2 PROJET SENSUNIQUE 19
2.2 ÉCRITS SCIENTIFIQUES 23
2.2.1 PROJET SARS 23
2.2.2 PLATEFORME D'APPRENTISSAGE DU LEXIQUE SPECIALISE 24

3. APPROCHES NORMATIVES EN REDACTION TECHNIQUE ET CONCEPT DES LANGUES CONTROLEES 25

3.1 ÉTAT DE L'ART SUR LES LANGUES CONTROLEES ET OUTILS D'AIDE A LEUR ECRITURE 26
3.2 PARAMETRAGE D'UNE LANGUE CONTROLEE 28
3.3 INFORMATISATION DU PROCESSUS D'ECRITURE EN LANGUE CONTROLEE : GENERATION DE TEXTES EN LC 33
3.3.1 COMPAGNON LISE : UNE INTERFACE COLLABORATIVE D'AIDE A LA REDACTION EN LC 34
3.3.2 AVANTAGES, INCONVENIENTS ET PERSPECTIVES POUR LE COMPAGNON LISE 38
3.4 LANGUES CONTROLEES SUR MESURE : LA PROBLEMATIQUE DU LEXIQUE 40
3.4.1 LEXIQUE D'UNE LANGUE CONTROLEE ET VOCABULAIRE CONTROLE 42
3.4.2 STRUCTURES LEXICALES 44
3.4.3 INFORMATISATION DU PROCESSUS DU RECENSEMENT DU LEXIQUE D'UNE LANGUE CONTROLEE : LA STATION SENSUNIQUE 46
3.4.4 CONCLUSION 71

4. REDACTION DE TEXTES SCIENTIFIQUES ET LEXIQUE SPECIALISE CONTEXTUALISE..... 74

4.1 SYSTEME D'AIDE A LA REDACTION SCIENTIFIQUE (SARS) DANS LE DOMAINE BIOMEDICAL..... 75
4.1.1 ÉTAT DE L'ART DES AIDES (INFORMATISEES OU NON) A LA REDACTION SCIENTIFIQUE..... 76
4.1.2 ENQUETE SUR LES HABITUDES REDACTIONNELLES 89
4.1.3 PRINCIPES DE CONCEPTION 95
4.1.4 PREMIERS RESULTATS DE NOS RECHERCHES 98
4.2 PLATEFORME D'APPRENTISSAGE DU LEXIQUE SPECIALISE 105
4.2.1 ÉTAT DE L'ART DE LA GENERATION AUTOMATIQUE D'EXERCICES 106
4.2.2 PREMIERES EXPERIMENTATIONS SUR LA GENERATION DES EXERCICES EN LANGUES SPECIALISEES 108
4.2.3 CONCLUSION..... 113

5. CONCLUSION ET PERSPECTIVES 115

5.1 PROJET TRADARC **117**
5.2 PROJET MENUSA.....**118**
5.3 ET A LONG TERME… **119**

6. REFERENCES BIBLIOGRAPHIQUES 121

7. ANNEXES 133

7.1 EXEMPLE D'UN PROTOCOLE CONTROLE : CHRU, NETTOYAGE DE CHAMBRES D'HOSPITALISATION (MEDECINE NUCLEAIRE) 133
7.2 STATION SENSUNIQUE : ALGORITHME DE PONDERATION.....134
7.3 STATION SENSUNIQUE : FORMAT D'EXPORT 136
7.4 100 FAMILLES LES PLUS FREQUENTES DANS LE LEXIQUE TRANS-BIOMEDICAL.....137

I. LISTE DES FIGURES

Figure 1 Schéma du projet Sensunique 21
Figure 2 Paramètres d'établissement de LC LiSe 29
Figure 3 Structure de la phrase en LC LiSe 35
Figure 4 Interface de 'Compagnon LiSe' avec 4 espaces de travail 36
Figure 5 Compagnon LiSe : Affichage des structures associées à un titre. 36
Figure 6 Compagnon LiSe : Exemples de structures verbales. 37
Figure 7 Compagnon LiSe : Structure désambigüisant d'un verbe 38
Figure 8 Compagnon LiSe : Définition associée à une structure verbale 38
Figure 9 Compagnon LiSe : Structures désambiguïsant du verbe 'assister' 39
Figure 10 Compagnon LiSe : Ecriture prédictive et étiquetage semi-intuitive 39
Figure 11 Compagnon LiSe : Choix d'une préposition pour l'introduction d'un complément circonstanciel du lieu 40
Figure 12 Schéma de la Station Sensunique du point de vue de ses fonctionnalités. 47
Figure 13 Interface de connexion de la Station Sensunique 54
Figure 14 Schéma de la Station Sensunique du point de vue chronologique 56
Figure 15 Station Sensunique : Interface du projet (onglet Projet) 57
Figure 16 Station Sensunique, paramétrages des pondérations 62
Figure 17 Station Sensunique : Statistiques de l'analyse 62
Figure 18 Station Sensunique : Interface de travail 65
Figure 19 Station Sensunique : Concordancier évolué 65
Figure 20 Station Sensunique : Liste des UL 66
Figure 21 Station Sensunique : Fiche lexicale d'une UL 67
Figure 22 Station Sensunique : Fiche de relation d'une UL 68
Figure 23 Station Sensunique : Visualisation des UL 69
Figure 24 Station Sensunique : Création d'une SL 70
Figure 25 Station Sensunique : Validation d'une UL/SL 70
Figure 26 Station Sensunique : Export du dictionnaire d'UL 71
Figure 27 Amadeus : Choix de structures rhétoriques 80
Figure 28 COBWEB : Interface d'édition pour les essais randomisés contrôlés 81
Figure 29 ARTES : Terminologie en contexte 82
Figure 30 ARTES : Phraséologie discursive 83
Figure 31 ARTES : Visualisation en corpus 83
Figure 32 Scientext : Recherche libre guidée 85
Figure 33 Linguee : Interface de recherche 86
Figure 34 TradooIT : Interface de recherche 87
Figure 35 SWAN : Évaluation d'une introduction 88
Figure 36 SWAN : Évaluation visuelle d'une introduction 89
Figure 37 Résultats d'enquête : Difficultés à utiliser l'anglais scientifique général 91
Figure 38 Résultats d'enquête : Difficultés à utiliser l'anglais scientifique médical 91
Figure 39 Résultats d'enquête : Difficultés à trouver des équivalents terminologiques 91
Figure 40 Résultats d'enquête : Termes simples et complexes 92

II. LISTE DES TABLEAUX

Tableau 1 Tâche de recensement des UL 51
Tableau 2 Tâche de recensement des termes 52
Tableau 3 Paramètres de pondération du PT 60
Tableau 4 Paramétrage de pondérations du PSL 61
Tableau 5 Paramétrage de pondérations du PUL 61
Tableau 6 Résultats d'enquête : Utilisation des dictionnaires papier 93
Tableau 7 Résultats d'enquête : Utilisation des dictionnaires en ligne 93
Tableau 8 Résultats d'enquête : Utilisation des manuels 93
Tableau 9 Résultats d''enquête : Utilisation des moteurs de recherche 94
Tableau 10 Résultats d'enquête : Outils à développer 94
Tableau 11 Premiers mots dans l'AWL (COXHEAD 2000) 100
Tableau 12 Premiers mots de la MAWL (WANG et al. 2008) 100
Tableau 13 MAWL : 27 familles de mots supprimées par les experts 101
Tableau 14 SARS : Description des journaux dans le corpus 104
Tableau 15 Projection des termes des listes sur le corpus support 110
Tableau 16 Projection des termes de la liste A ∪ B sur le corpus de référence 111
Tableau 17 Exemples d'indices 112
Tableau 18 Exemples de catégories sémantiques spécifiques au domaine de spécialité 114

III. LISTE DES ABREVIATIONS

ArgX	Argument X
MAWL	Medical Academic Word List
AWL	Academic Word List
CA	Corpus d'Analyse
CC	Corpus Contrastif
CHRU	Centre Régional Hospitalier Universitaire
CS	Corpus Support
CT	Candidat(s) Terme(s)
EdT	Extracteurs de Termes
EFS B/FC	Établissement Français du Sang Bourgogne / Franche Comté
Gn_sa	Groupe nominal sans article
LC	Langue(s) Contrôlée(s)
LLC	Lexique d'une Langue Contrôlée
neg	négation
opt	optionnel
PSL	Poids de Structure Lexicale
PT	Poids Terminologique
PUL	Poids d'Unité Lexicale
SARS	Système d'Aide à la Rédaction Scientifique
SL	Structure(s) Lexicale(s)
TAL	Traitement Automatique des Langues
UL	Unité(s) Lexicale(s)
ULC	Unité(s) Lexicale(s) Candidate(s)
Vconj	Verbe conjugué
Vinf	Verbe à l'infinitif

1. INTRODUCTION

J'ai compris que tout le malheur des hommes venait de ce qu'ils ne tenaient pas un langage clair.

Albert Camus

1.1 Textes spécialisés et contextes de rédaction

Lorsque je parcours l'ensemble de mes travaux universitaires, j'y décèle plusieurs invariants. En premier lieu, il y est souvent question de **textes écrits en langues spécialisées**. Ceux sur lesquels j'ai travaillé à travers mes différents projets, peuvent être subdivisés en deux grandes catégories. La première est constituée par des documents techniques, produits par des institutions à l'intention de leurs agents ou à l'intention de leurs usagers non-spécialistes. Ceci peut être illustré par des modes opératoires de l'Établissement Français du Sang Bourgogne / Franche Comté (EFS B/FC), destinés aux techniciens du laboratoire de l'immunobiologie, ou par les recommandations que l'EFS B/FC émet en direction de donneurs de sang bénévoles. Le second type de documents concerne les écrits scientifiques, plus précisément les articles que les chercheurs produisent pour communiquer leurs résultats à l'ensemble de leur communauté scientifique.

Une des caractéristiques communes aux textes spécialisés est qu'ils sont produits par des experts. Ce sont certes des spécialistes de leur domaine, mais ils ne sont pas nécessairement formés pour la rédaction[1]. Ce fait se reflète souvent dans la qualité des documents produits : on y trouve des problèmes de logique, de cohérence, de compréhension, des formulations ambigües ou floues. En fonction des contextes et des objectifs de la rédaction des textes, ces problèmes peuvent s'avérer importants, cruciaux, voire vitaux. La qualité du texte est importante pour un étudiant en géographie de l'eau qui produit son premier mémoire de recherche et espère une bonne notation. Elle l'est aussi pour un chercheur soumettant un article en anglais à une revue renommée qui exige une excellente qualité d'écriture. Enfin, la qualité du texte est vitale pour quelqu'un qui doit se servir d'un défibrillateur afin d'analyser l'activité du cœur d'une personne en arrêt cardio-respiratoire. Et on peut parier que la grande majorité d'entre nous ne lisent pas la notice d'utilisation d'un défibrillateur avant de n'y être contraint ! La qualité du texte devrait donc être (et l'est dans la majorité des cas) une préoccupation constante de ses concepteurs.

Ce premier invariant définit bien mon objet d'étude : ce sont les **textes spécialisés, dans le contexte particulier de leur production**. Ce contexte peut être analysé comme n'importe quelle situation de communication : l'objet y est produit en fonction des interlocuteurs, des circonstances de communication, des objectifs et des intentions des intervenants. Pour produire un texte, il est éminemment important de tenir compte du contexte de son écriture, parce que « *(…) la description d'une langue ne peut se limiter à établir*

[1] Pour les chiffres détaillés sur les experts sollicités par les tâches de rédaction, voir REJEAN (2000).

son système de règles ou d'unités » comme le dit très justement Maria Térésa CABRE (1998 :115).

Les contextes de rédaction en langues spécialisées, très variés, possèdent des caractéristiques particulières qui ne relèvent pas toujours à strictement parler de la linguistique. Pourtant, ils imposent des contraintes sur la manière d'appréhender l'écriture et par conséquent, sur les méthodologies et outils que l'on puisse développer pour aider les rédacteurs. Par exemple, lorsque l'exactitude et l'efficacité de la transmission de l'information sont vitales, les textes produits (en amont) devraient minimiser le risque d'erreurs (en aval) : c'est cette constatation qui nous a amenés à proposer la méthodologie d'écriture en Langue Contrôlée pour tout texte protocolaire[2] relevant des secteurs à risque et décrivant une suite d'actions à effectuer pour atteindre un objectif. Les avantages de cette méthodologie sont nombreux : elle permet d'obtenir des textes techniques clairs, structurés et exempts d'ambigüité, conformes à des modèles et des normes établis. Mais elle influe considérablement sur la façon de procéder du rédacteur, qui sera confronté, par exemple, au défi de la mémorisation : puis-je utiliser une telle structure syntaxique ? Puis-je utiliser ce mot sous cette acception ? Comment avais-je appelé ce concept auparavant ? Chaque contexte de rédaction apporte ses propres défis : un étudiant en médecine écrivant un article scientifique en anglais n'aura pas nécessairement besoin des explications linguistiques détaillées sur la terminologie à utiliser, alors qu'un linguiste concevant une langue contrôlée devra encoder toutes les informations sur les lexies spécialisées à recenser.

C'est précisément ces questionnements autour de la façon d'aider un rédacteur à construire un texte dans un contexte particulier de rédaction qui guident les travaux décrits dans ce mémoire.

1.2 Langues spécialisées et linguistique

Le second invariant de mes travaux concerne leur **ancrage dans la linguistique**. Les rapports entre la linguistique et les langues spécialisées n'ont pas toujours fait l'unanimité et restent encore source de polémiques et d'approches contradictoires. Il y a ceux qui réclament une langue de spécialité complètement distincte de la langue générale (HOFFMANN 1979) et ceux qui la considèrent comme une simple variante de la langue générale (RONDEAU 1983, QUEMADA 1978). D'aucuns la rangent comme un sous-ensemble dans la langue générale (CABRE 1998) et d'autres lui confèrent le statut de langue à part entière (LERAT 1995). Enfin, il y a ceux qui la définissent en termes avant tout linguistiques (L'HOMME 2004), et ceux qui clament l'impossibilité de la définir en termes purement linguistiques (PICHT et DRASKAU 1985, SAGER et al. 1980).

À cela s'ajoutent les discussions sur la terminologie et sur son autonomie par rapport au langage, surtout depuis l'émergence, dans les années 90, du courant de la terminologie

[2] Un protocole est un texte qui vise à communiquer à un utilisateur, qu'il soit ou non spécialiste du domaine, les actions à accomplir pour attendre un objectif, sous certaines conditions. Un protocole peut être destiné à une exécution immédiate, avec des différents niveaux d'urgence, ou il peut instruire un utilisateur potentiel sur les actions à entreprendre en cas d'urgence (cf. &3.2).

textuelle, représentée par les chercheurs tels que Maria Térésa CABRE en Espagne, Didier BOURIGAULT et Monique SLODZIAN en France ou Marie-Claude L'HOMME au Canada.

Pour ma part, j'aborde la langue spécialisée en m'appuyant sur la définition de LERAT (1995 : 20) qui dit qu'elle est « *une langue naturelle considérée en tant que vecteur de connaissances spécialisées* ». Je l'aborde par le biais de textes — et plus particulièrement par le biais de la production de textes spécialisés — qui nécessite une mobilisation des compétences linguistiques de la part des rédacteurs, en plus des compétences liées au domaine. Lorsque l'on pose aux experts la question de ce qui leur pose le plus de difficultés dans la rédaction des textes spécialisés en langue étrangère, par exemple, ce n'est pas tellement la connaissance de termes liés à leur domaine, mais bel et bien leur mise en contexte[3] dans un écrit. Ils possèdent tout naturellement les connaissances propres à leurs domaines, alors qu'ils éprouvent des difficultés à les exprimer dans un texte.

Je prends aussi soin de distinguer, du moins de façon théorique, les *termes* des *lexies spécialisées*. Cette distinction est souvent difficile à maintenir dans le langage, tellement la notion de *terme* s'est imposée dans l'usage pour designer tout mot référant à un domaine de spécialité. La distinction que je fais entre ces deux notions dans mon travail provient de son orientation vers le texte et l'écriture et est de nature fonctionnelle. Les termes d'un domaine sont établis dans un objectif spécifique : conceptualisation d'un domaine, à travers un processus particulier, impliquant une normalisation et une validation par des comités des experts : ses principes méthodologiques sont définis depuis les années 1950 au niveau international par le Comité technique 37 de l'ISO (ISO TC37) chargé d'élaborer et de normaliser les procédures de travail en matière de terminologie[4]. Alors que les lexies spécialisées servent à construire un discours spécialisé, donc produire des textes spécialisés, et ceci en dépit du fait qu'elles soient ou non validées par une norme.

Serait-il souhaitable, et si oui, serait-il possible de construire les textes spécialisés en n'utilisant que des termes officiellement approuvés ? C'est une question que nous nous sommes évidemment posée lors de nos travaux sur les langues contrôlées. Des discussions très intéressantes entre les partisans de différentes théories de la terminologie plaident en faveur de telle ou telle définition du terme et du processus de son établissement[5]. Les défenseurs des théories issues de la tradition wüstérienne plaident l'impossibilité de superposer la structure informationnelle d'un discours — d'ordre linguistique — et la structure conceptuelle du monde, qui serait d'ordre scientifique (DEPECKER et ROCHE 2010). Les représentants des courants textuels en terminologie pointent l'idéalisme, l'insuffisance et l'artificiel de la théorie conceptuelle en lui opposant les approches empiristes, tenant compte de la variabilité en contexte et de connaissances pertinentes inscrites dans les textes (SLODZIAN 2000). Pour ma part, je me positionne délibérément du côté du texte et du fait linguistique que je peux y observer ou en déduire, sans intervenir sur le terrain des compétences du domaine, nécessaire à sa conceptualisation. D'où ma préférence pour utiliser

[3] Voir les résultats d'enquête auprès des professionnels de la Santé (cf. &4.1.2).
[4] Afnor, www.afnor.fr.
[5] Voir à ce sujet l'excellent article de John HUMBLEY (2004), *La réception de l 'œuvre d'Eugen Wüster dans les pays de langue française*.

la notion de *lexie spécialisée* qui renvoie immédiatement vers l'analyse lexicale d'un texte. Et même avec ce programme volontairement linguistique, il est souvent impossible de se passer de l'expert pour, par exemple, construire un dictionnaire des lexies spécialisées à partir des textes. Ainsi la pratique des langues spécialisées impose une transdisciplinarité, qui, à elle seule, garantit une description adéquate de l'objet d'étude.

Le domaine où se rencontrent certains préceptes de la terminologie dans son acception conceptuelle et de la linguistique est celui de langues contrôlées, auxquelles on reproche d'ailleurs les mêmes choses qu'à la terminologie conceptuelle. Une langue contrôlée ne serait pas viable puisqu'elle est une sorte d'idéalisation sur le fonctionnement de la langue et de ce fait, elle serait impossible à mettre en pratique. Une langue contrôlée exige une très forte normalisation, alors que les textes témoignent de la diversité de l'emploi de formes et de sens. Une langue contrôlée ne serait pas suffisante pour exprimer toutes les nuances du contenu d'un texte, puisqu'elle ne permet qu'un nombre limité de sens et de structures. Elle est aussi artificielle, dans le sens où ses frontières sont découpées arbitrairement dans l'ensemble des règles d'une langue naturelle. Pourtant, en termes de besoins communicationnels, ce sont les langues contrôlées qui répondent le mieux aux principes de clarté, simplicité et lisibilité, nécessaires à la bonne transmission de l'information spécialisée. L'établissement de langues contrôlées relève d'une démarche normative, voire prescriptive, tout comme l'établissement de la terminologie selon la théorie conceptuelle. C'est la démarche normative et prescriptive qui permet d'atteindre les objectifs que l'on impose à la communication spécialisée.

1.3 Lexiques spécialisés et contextualisation

L'ancrage de mes travaux dans la linguistique se manifeste aussi par l'intérêt particulier que je porte au **lexique des langues spécialisées**. Selon CABRE (1998 : 133) le lexique est « le niveau le plus important » dans les écrits basés sur les langues spécialisées. RONDEAU (1983) surenchérit sur le fait que les langues spécialisées se caractérisent fondamentalement par leur lexique. Derrière cette notion, très large, du lexique d'une langue spécialisée se dessinent plusieurs problématiques, relevant aussi bien de sa description théorique que du besoin de le manipuler informatiquement et de le rendre accessible à un utilisateur, à savoir : qu'est-ce qui compose le lexique de langues spécialisées ? Comment reconnaitre et recenser le lexique des langues spécialisées dans les textes? Comment le formaliser pour les systèmes automatisés ? Comment le décrire pour un utilisateur non-linguiste ?

Le lexique d'une langue spécialisée ne se réduit pas au lexique spécifique du domaine. CABRE (1998 : 137) propose de distinguer 3 types de vocabulaires, selon le dégrée de leur spécialisation ('commun', 'intermédiaire' et 'spécialisé') et elle y ajoute la présence des codes provenant des autres systèmes sémiotiques, telles que la chimie et les mathématiques[6]. CAMLONG (1996) va jusqu'à distinguer 8 types de vocabulaires, constituant un continuum allant du vocabulaire terminologique du domaine au vocabulaire général. Pour notre part, nous avons conceptualisé le type du lexique à décrire en fonction du contexte de rédaction.

[6] Voir à ce sujet les articles d'Yves GENTILHOMME (cf. bibliographie).

Par exemple, dans le contexte d'aide à l'écriture en langues contrôlées, nous avons établi le concept du *Lexique d'une Langue Contrôlé*e (cf. &3.4.1), pour répondre aux exigences de non-ambigüité et de non-redondance, c'est-à-dire au fait que, dans une Langue Contrôlée, une unité lexicale ne peut avoir qu'une seule définition, et qu'une définition ne peut correspondre qu'à une seule unité lexicale dans un domaine choisi. Pour ce faire, il s'est avéré nécessaire de contrôler l'ensemble du lexique utilisé pour la conception de la documentation dans un domaine, ainsi que recueillir les informations complémentaires concernant les liens qu'entretiennent les différentes unités entre elles (synonymie, inclusion, variation etc.). Ainsi pour mettre en place une démarche normative et prescriptive, il faut d'abord décrire l'existant dans la langue. Dans le contexte de l'aide à l'écriture des articles scientifiques en biomédical, nous avons proposé le concept de *lexique trans-biomédical* (voir &4.1.4.1), qui, en s'appuyant sur l'idée du lexique scientifique transdisciplinaire, est défini comme un lexique constitué des lexies spécialisées communes à l'ensemble des sous-domaines biomédicaux.

Quel que soit le lexique que nous construisons (lexique d'un domaine, lexique trans-biomédical, etc.) nous cherchons à le décrire en contexte, constituant ainsi des **lexiques spécialisés contextualisés**. Les lexies spécialisées sont caractérisées par « *une syntagmatique restreinte (cooccurrences et commutations dans les limites d'un domaine spécialisé)* » (LERAT 1995 : 52), il est donc possible d'énumérer leurs contextes significatifs. Le choix de contextualisation est primordial pour aider un rédacteur en langue contrôlée, à qui on souhaiterait suggérer la suite de sa saisie en fonction du texte déjà encodé. L'apprentissage du lexique en contexte guide la construction de du système d'aide à l'apprentissage des langues spécialisées. Enfin, c'est trouver le mot exact pour accompagner un terme dans un texte en langue étrangère qui pose le plus de difficultés aux médecins - rédacteurs des articles dans leur spécialité. Les travaux sur les collocations en langues spécialisées et sur les avantages que la description des collocations spécialisées peut apporter aux rédacteurs d'articles témoignent de l'importance de ce sujet (VOLANSCHI et KÜBLER 2010 ; KÜBLER et PECMAN 2012 ; L'HOMME 2013). Le choix de contextualisation du lexique spécialisé est aussi inspiré par des travaux concernant le lexique scientifique transdisciplinaire, qui donnent autant d'importance à l'étude du lexique qu'à l'analyse et le recensement de la phraséologie scientifique (DROUIN 2007, TUTIN 2007). Nous avons aussi exploré du côté des études sur le recensement des cooccurrents dans les textes et sur la description des propriétés collocatives des unités lexicales dans les dictionnaires d'apprentissage, surtout dans les dictionnaires d'un type un peu nouveau comme DiCoInfo (L'HOMME 2008) ou le DAFA, Le Dictionnaire d'Apprentissage du Français des Affaires, développé à l'Université de Leuven en Belgique (BINON et al. 1992).

1.4 Vers les besoins des utilisateurs

L'objectif applicatif affirmé de mes travaux est la construction d'outils d'aide à l'écriture en langues spécialisées. Il en résulte le troisième invariant, à savoir l'orientation sur les **besoins des utilisateurs**.

Dans cette optique centrée utilisateur, le **rôle de l'expert** du domaine est essentiel. On le retrouve en plusieurs positions. Premièrement, il est spécialiste de son domaine, il est celui

qui peut établir et préciser les concepts et les notions. Deuxièmement, il est souvent le rédacteur des textes spécialisés, peu entrainé à le faire, donc peu conscient des problématiques langagières dont ceux-ci sont porteurs. Il ne maitrise pas bien le jargon des linguistes, il est donc préférable d'encoder l'information linguistique de façon simple, en prenant soin de ne pas alourdir inutilement ni la quantité des données ni la façon dont elles sont présentées. Le troisième rôle de l'expert est celui d'un « maître d'ouvrage », à savoir celui qui sait le mieux définir les besoins, parce qu'il représente souvent les utilisateurs finaux à qui l' « ouvrage » est destiné ; de ce fait, son implication est indispensable lors de la réflexion sur les besoins. Finalement, l'expert est celui qui valide : que ce soit au stade intermédiaire d'une recherche (par exemple, le lexique à retenir pour la construction des langues contrôlées) ou au stade final (le système d'aide à l'écriture en langues contrôlées). Comme dit Pierre LERAT (1995 : 47) en ce qui concerne les langues spécialisées, '*en fin de compte, le critère de critère est l'avis de spécialiste, dont on ne saurait raisonnablement tenter de faire l'économie*'.

Le linguiste profite de la collaboration avec un expert du domaine, mais l'expert, lui aussi, tire des avantages de sa collaboration avec le linguiste. En essayant de désambigüiser et de 'mettre au clair' sa langue, le linguiste, avec son regard de 'novice' permet à l'expert de réfléchir sur sa pratique métier, sur la façon dont il l'a décrite et sur son processus rédactionnel. Cette réflexion nous a souvent été faite durant nos projets par les spécialistes de différents domaines avec qui nous avons travaillé. Par ailleurs, il est intéressant de faire le rapprochement entre les recommandations linguistiques données aux auteurs des articles scientifiques et les règles selon lesquelles on construit les langues contrôlées : les deux ont pour objectif de clarifier le langage, et à travers le langage, l'ensemble des pratiques soumises à la description. À mon avis, ce serait une des fonctions très importantes de la langue dans le contexte de l'écriture en langues spécialisées : elle permettrait de clarifier les pratiques métier.

1.5 Outiller la rédaction

Le quatrième invariant de mes travaux concerne les caractéristiques des outils que nous construisons.

Premièrement, leur conception s'appuie sur les modèles linguistiques établis en amont à partir de l'analyse des corpus. Pour certaines tâches, orientées expert, nous avons pris le soin d'utiliser des corpus relativement petits (quelques dizaines de milliers de mots), même si la tendance aujourd'hui est de compiler des corpus de plus en plus volumineux. Ceci pour deux raisons principalement. D'une part, la construction d'un corpus pour chaque nouveau projet impliquant la terminologie n'est pas une tâche triviale, surtout pour un expert. Par conséquent, plutôt que d'investir dans la quantité des données, nous avons fait le choix d'améliorer le processus de filtrage des propositions que font des outils automatiques pour faciliter la prise de décision par un expert. D'autre part, nous nous sommes rendu compte que certaines institutions ne disposent pas de grandes collections de textes : ceci était le cas de l'Établissement Français du Sang, par exemple, qui nous a transmis l'ensemble de ces modes opératoires concernant l'activité d'immunobiologie durant le projet Sensunique. Ceci est aussi confirmé dans les travaux de Patrick DROUIN (2002) qui, pour tester son extracteur de termes,

a utilisé un corpus comparable aux nôtres, provenant d'une entreprise privée et décrit comme représentatif de leur fond documentaire.

Deuxièmement, tous nos outils sont destinés à un utilisateur humain ; par conséquent, leurs résultats doivent être appréhendables et, si nécessaire, modifiables par un être humain. Ceci est très important, par exemple, lorsque l'on conçoit un outil d'aide au recensement et à la gestion du lexique en langue contrôlée : même si on utilise les automatismes comme aide à la décision terminologique, on laisse toujours le dernier mot au linguiste. De même, lorsqu'on propose à un professeur en langues spécialisées la construction d'un nouvel exercice d'apprentissage du lexique, on lui laisse le choix de ses exemples parmi toutes les propositions faites par l'outil.

Cette orientation résulte principalement de deux constats. Le premier concerne la qualité des textes spécialisés qu'il nous faut obtenir : elle est indispensable pour les domaines applicatifs choisis dans nos projets (dits domaines à haut risque, tels que la médecine nucléaire dans le projet Sensunique ou l'aéronautique dans le projet LiSe), mais résulte aussi de types de textes que nous voulons aider à produire, par exemple les articles à publier dans les revues scientifiques. Pour ces types de tâches, les systèmes automatiques ne peuvent pas produire les résultats entièrement fiables même si, par ailleurs, leurs résultats sont suffisants pour certaines autres applications. Par exemple, la maturité des extracteurs terminologiques est aujourd'hui certainement suffisante pour construire les vocabulaires contrôlés dans le contexte de la recherche d'information. Cependant, lorsqu'il s'agit de construire de dictionnaires spécialisés ou de traduire, les extracteurs produisent (et produiront toujours) trop de bruit pour que leurs résultats puissent être acceptés sans une supervision humaine.

Troisièmement, nous nous sommes rendu compte d'un besoin croissant de spécialisation des outils. La spécialisation peut signifier l'adaptabilité à un domaine, voire à un sous-domaine, voire à une thématique attachée à un sous-domaine, mais aussi l'adaptabilité à un projet particulier ou aux besoins d'une institution. C'est de ce constat que provient le concept d'une langue contrôlée 'sur mesure', à savoir une langue contrôlée qui répond aux besoins d'une structure, d'une institution ou d'un établissement en particulier. Ce serait aussi une demande des médecins d'un outil d'aide à la rédaction des articles : ils désireraient avoir des outils spécifiques pour la cardiologie, pour la gastro-entérologie, pour les essais cliniques et pour les essais randomisés, pour n'en citer que quelques parmi de nombreux exemples. De ce fait, nous donnons une importance particulière aux méthodologies généralisables et à la construction d'outils adaptables aux domaines et paramétrables suivant les besoins de l'utilisateur. La méthodologie mise en place pour la création des exercices en langues spécialisées pour la géographie de l'eau doit, par exemple, être opérante pour tout autre domaine pour lequel un enseignant de langues spécialisées souhaiterait l'appliquer.

De façon générale, les outils sont toujours les résultats d'un savant compromis entre la modélisation d'une problématique, les besoins des utilisateurs et les contraintes des systèmes automatisés. Par exemple, la modélisation des structures syntaxiques en langue contrôlée a été étroitement liée à la problématique de leur apprentissage par un rédacteur potentiel, et de ce fait, à la façon dont une interface peut aider à leur écriture. La notion du lexique trans-biomédical contextualisé a émergé suite au besoin de construire des outils très

spécialisés (en occurrence par sous-domaine biomédical), mais qui auraient néanmoins une partie de leurs lexiques en commun.

L'aspect applicatif de notre travail est aussi le résultat du contexte dans lequel il s'est déroulé. Pour travailler sur les textes de spécialité, nous avons souhaité mettre en place un contexte transdisciplinaire, pour pouvoir appuyer nos choix sur ceux des experts de domaine. L'orientation applicative permet un travail interdisciplinaire équilibré : pour ne prendre qu'un exemple, même si les chercheurs en immunologie perçoivent l'intérêt de travailler sur leur langue de spécialité, ceci ne constitue pas leur préoccupation de recherche, puisque la langue n'est pour eux qu'un outil pour exprimer leurs connaissances. De ce fait, une recherche fondamentale sur les langues spécialisées risque de ne pas être assez 'attractive' pour eux. En revanche, une orientation plus appliquée permet de concrétiser les enjeux et d'avoir des retombées tangibles en retour (par exemple, les textes réécrits en langues contrôlées). Ceci est un argument très valable pour les financeurs de la recherche aussi. Par conséquent, les chercheurs impliqués dans les projets applicatifs usent des fois de leur liberté pour mener conjointement leur recherche en direction de l'application et une recherche plus fondamentale sur le même sujet. Finalement, pour faire avancer les recherches en TAL, il est nécessaire de disposer d'outils 'intermédiaires', dont les résultats sont indispensables pour valider des hypothèses de recherche plus avancées. C'était par exemple l'idée derrière le développement de la Station Sensunique : au-delà de son but premier qui consistait à aider la construction des lexiques de langues contrôlées, la plateforme devait disposer de plusieurs configurations lui permettant de fonctionner dans de nombreux contextes d'utilisation, certains complétement automatisés.

1.6 Vers l'autonomie des utilisateurs

L'objectif final de mes travaux concerne l'**autonomie** que l'outil devrait donner à l'utilisateur par rapport à la tâche qu'il veut accomplir. On peut définir l'autonomie en termes de capacité de l'utilisateur à accomplir l'action sans avoir recours à d'autres supports extérieurs. Ceci ne signifie pas la même chose pour tous les types d'utilisateurs, ni pour tous les outils. Lorsqu'on a créé le Compagnon LiSe, l'interface d'aide à l'écriture en langues contrôlées, on l'a destiné à un utilisateur non-linguiste, de surcroît inconscient de la complexité que ce type d'écriture peut impliquer. Le rôle du Compagnon LiSe consiste à garantir une rédaction de la part de l'utilisateur régie des principes d'une langue contrôlée, tout en veillant à ce qu'il ne soit pas bloqué par l'impossibilité de formuler des phrases. Cela suppose que la manière dont l'outil va guider l'utilisateur vers la rédaction des protocoles et de sa capacité à anticiper les difficultés (par exemple, le choix de structures syntaxiques autorisées ou le remplacement de termes interdits) sont d'une importance capitale. Pour ce type d'utilisateurs, l'autonomie dans la rédaction résulte d'une aide très soutenue par le logiciel qui les empêche de se mettre dans une situation d'impossibilité d'accomplir l'action. D'un autre côté, la Station Sensunique qui est *a priori* destinée à des linguistes, utilise des automatismes pour produire de l'information, mais laisse à l'analyste la liberté totale de les accepter, de les rejeter ou de les modifier. L'utilisateur ne doit pas être non plus dépendant de l'outil, qui doit s'adapter à ses besoins, tout en lui procurant la meilleure aide possible.

Voilà quelques constantes qui se dégagent de mes travaux. De surcroît, ces constantes situent mon parcours à la jonction de plusieurs domaines, tels que la lexicologie et la lexicographie (spécialisées), la rédaction technique, l'enseignement et l'apprentissage de langues spécialisées, la terminologie, la normalisation et le TAL. Elles constituent un cadre unificateur à l'ensemble des projets que j'ai coordonnés ou auxquels j'ai participé. Chaque projet a apporté un nouveau contexte de rédaction en langue spécialisée, donc une nouvelle problématique, soit par un nouveau type de documents à étudier, soit par un public diffèrent, soit par un objectif singulier à atteindre.

1.7 Organisation du mémoire

Pour rendre compte de ces problématiques, la suite de ce mémoire s'organise de la façon suivante. Je présente d'abord les enjeux, les problématiques et les objectifs des projets de recherche qui ont structuré ma vie de chercheuse (et qui, pour certains, continuent de le faire) (&2). Ils sont souvent plus larges que certains de leurs résultats présentés dans ce mémoire. Ces projets sont classifiés selon deux axes : ceux qui concernent les écrits professionnels (&2.1) et ceux qui ont pour objet d'étude les écrits scientifiques (&2.2). Le chapitre 3 est consacré aux méthodes et aux outils que nous avons développés pour soutenir l'écriture en langues spécialisées. Tout d'abord, je présente le concept de l'écriture en langues contrôlées, ses avantages et ses inconvénients, ainsi que l'interface que nous avons développées pour les contourner, le Compagnon LiSe (&3.3). La seconde partie de ce chapitre est consacrée à la conception et la gestion des lexiques qui permettent de concevoir les langues contrôlées 'sur mesure', et à la plateforme qui a été conçue expressément dans cet objectif, la Station Sensunique (&3.4). Le chapitre 4 aborde la rédaction de textes scientifiques et le rôle des lexiques spécialisés contextualisés, sous deux angles : d'abord à travers un système d'aide à l'écriture de textes scientifiques en biomédical (&4.1) et ensuite à travers le projet d'une plateforme d'apprentissage de ces lexiques destinée aux étudiants de l'université (&4.2). La conclusion et les perspectives de recherche finissent le tour d'horizon de mes travaux (&5).

2. CONTEXTES DE REDACTION EN LANGUES SPECIALISEES

On joue d'un objet exceptionnel, dont la linguistique a bien souligné le paradoxe : immuablement structuré et cependant infiniment renouvelable : quelque chose comme le jeu d'échecs.

Roland Barthes

Les textes spécialisés sont porteurs de caractéristiques qui les distinguent de tout autre type de textes et qui leur confèrent une certaine unité. Tout d'abord, ils portent sur un sujet particulier, lié à un domaine de connaissances qui n'est pas familier à tous les utilisateurs d'une langue. Leur fonction fondamentale est de transmettre l'information, de façon précise et sans ambigüité, ce qui influe sur l'absence (ou du moins sur l'atténuation) des autres fonctions du langage présentes. Ils sont impersonnels, tentent à dissimuler la présence de l'auteur, ne cherchent pas à embellir la langue. Ils sont écrits selon une structure préétablie et possèdent des caractéristiques particulières en termes de leurs composantes morphologiques, syntaxiques et, surtout, lexicales. Ils sont hautement codifiés, ce qui plaide en faveur de leur traitement informatisé. On n'écrit pas des textes spécialisés de manière spontanée : leur production exige un apprentissage des règles qui régissent leur forme textuelle, les structures et le lexique que l'on va utiliser.

À côté de ces caractéristiques unificatrices, les textes de spécialité se différencient entre eux selon plusieurs paramètres : certains résultent des propriétés intrinsèques des textes, mais d'autres proviennent des contextes de leur production. Ces contextes peuvent être analysés comme n'importe quelle situation de communication : le texte y est produit en fonction des interlocuteurs, des circonstances de communication, des objectifs et des intentions des intervenants. Pour appuyer ce fait, CABRE (1998 : 121) parle de deux types de spécialisation :

- spécialisation par le sujet : c'est ce type de spécialisation qui est le plus souvent invoqué pour caractériser les langues spécialisées ;
- spécialisation par les caractéristiques « particulières » de l'échange d'informations ; bien que ces caractéristiques ne relèvent pas, à strictement parler, de la linguistique, elles imposent des contraintes sur la manière d'appréhender l'écriture des textes.

Les critères de caractérisation des contextes de rédaction sont les suivants :

a) le domaine ;
b) les types des textes spécialisés ;
c) le type des rédacteurs ;
d) le public ;
e) l'objectif de la communication.

Dans cette section, nous allons présenter les différents projets qui ont structuré notre recherche autour des langues spécialisées. Chacun de ces projets définit un contexte de

rédaction différent, ce qui nécessite une approche particulière et une aide informatisée spécifique. En ce qui concerne les sujets et les domaines, nous nous sommes beaucoup appuyés sur les différents sous-domaines du biomédical (immunobiologie, médecine nucléaire), grâce à la collaboration que nous avons développée avec les structures de recherche dans le domaine du biomédical de l'Université de Franche-Comté (SFR FED 4234), avec le CRHU (Centre Régional Hospitalier Universitaire) de Besançon et avec l'Établissement Français du Sang Bourgogne/Franche-Comté (EFS B/FC). Au-delà des domaines biomédicaux, nous avons aussi travaillé dans les domaines de la sécurité civile et de la géographie de l'eau.

Les types de textes spécialisés que nous avons abordés varient en fonction des projets. Pour la rédaction technique, nous avons utilisé les protocoles et les modes opératoires provenant des institutions partenaires. Le corpus des textes scientifiques a été construit à partir des publications disponibles sur l'Internet. En ce qui concerne les rédacteurs, ce sont toujours les spécialistes de domaine : ceux que CABRE (1998 : 36) appelle les *usagers directs* pour les distinguer des *intermédiaires*, à savoir les professionnels de la langue : traducteurs, rédacteurs, terminologues, qui se servent de langues spécialisées pour faciliter la communication à d'autres usagers. Les publics, ceux à qui les textes spécialisés sont destinés, ne sont pas aussi homogènes : bien que la communication spécialisée soit souvent adressée à des spécialistes (ce qui est le cas de la communication scientifique, par exemple), ce n'est pas toujours le cas. Jean-Marie KLINKENBERG distingue dans ce cas la communication horizontale – celle qui se passe entre les pairs - et verticale – '*…lorsqu'elle fait passer les données entre partenaires ne disposant pas des mêmes informations ou du même niveau de formation*' (2000 : 23). Par exemple, certains protocoles concernant la gestion des déchets nucléaires sont destinés au personnel de nettoyage dans le service de Médecine Nucléaire à l'hôpital de Besançon, qui n'a pas de formation en médecine nucléaire. D'autres, du type consignes de sécurité ou attitudes à adapter en cas d'un cataclysme, par exemple, s'adressent au grand public. En effet, les objectifs de la communication par textes spécialisés ne sont pas les mêmes lorsqu'il s'agit de transmettre des instructions précises à des professionnels dans le cadre de leur travail, lorsqu'il s'agit de s'adresser au grand public lors d'une situation de crise ou encore lorsqu'il s'agit de communiquer auprès de scientifiques à travers les journaux spécialisés.

Nous avons regroupé ces contextes de rédaction en deux catégories : la première concerne les écrits professionnels, c'est-à-dire les écrits relevant de la rédaction technique dans une situation professionnelle (&2.1), la seconde concerne les écrits scientifiques (&2.2). Nous allons rapidement présenter les projets qui ont structurés nos travaux sur ces deux types d'écrits et qui ont conduit à la conception, et pour certains la réalisation, des outils d'aide à l'écriture en langues spécialisées.

2.1 Écrits professionnels (rédaction technique)

2.1.1 Projet LiSe

Le projet ANR LiSe (*Linguistique, normes, traitement automatique des langues et Sécurité: du Data et Sense Mining aux langues contrôlées* ; coordinateur et porteur du projet :

Sylviane Cardey)[7], qui s'est déroulé au Centre Tesnière entre 2007 et 2010 a été le premier projet français à mettre l'accent sur le concept de Langues Contrôlées (LC) et sur l'utilité des LC pour la production et la traduction de textes techniques de qualité. Le défi du projet LiSe consistait à créer une LC française, qui faciliterait l'écriture et la traduction de protocoles et alertes dans les domaines de haute sécurité (santé et aéronautique) de façon fiable et non-ambigüe. L'analyse de plusieurs corpus de textes authentiques a permis de définir les problèmes auxquels se heurtent les textes techniques, aussi bien en ce qui concerne leur qualité (CARDEY 2009; VUITTON et al. 2009; RENAHY et al. 2015) que les contraintes qui pèsent sur leur rédaction (RENAHY et al. 2012). Nous avons établi en conséquence des normes de contrôle pour le français en tenant compte de plusieurs contraintes (RENAHY et al. 2009):

- contraintes générales liées à l'utilisation d'une LC (lisibilité et non-ambigüité) : structuration d'information, contrôle syntaxique, contrôle lexical ;
- contraintes liées au type du document : structuration interne du document ;
- contraintes liées au domaine : niveau de sécurité exigé, normes rédactionnelles en usage, contraintes terminologiques spécifiques ;
- contraintes liées à la traduction automatique pour assurer la qualité de traduction vers les langues cibles du projet.

Pour faciliter la gestion de normes de contrôle par un rédacteur technique, nous avons créé une interface interactive d'aide à la rédaction de textes techniques, appelée Compagnon LiSe (RENAHY and THOMAS 2009). Ce logiciel automatise une grande partie de règles de contrôle et propose une aide interactive pour les règles qui ne peuvent pas être automatisées. Le rédacteur technique, qui *a priori* n'est pas un linguiste, rédige les protocoles en s'appuyant sur un système de structures fonctionnelles et syntaxiques préétablies ou en choisissant des options disponibles dans les menus déroulants.

Ce logiciel assure deux objectifs complémentaires :

- d'une part, il répond aux critiques souvent formulées à l'encontre des LC, à savoir la difficulté d'apprendre et de gérer l'ensemble de normes de contrôle, surtout par les non-linguistes ;
- d'autre part, il garantit la qualité rédactionnelle de textes techniques écrits selon les normes de contrôle, quel que soit le rédacteur.

2.1.2 Projet Sensunique

Le projet ANR Sensunique (*Sensunique : pour une rédaction optimale de textes techniques de qualité assistée par un rédacticiel innovant,* ANR-2010-EMMA-039, 2010-2012, porteur et coordinateur du projet : Izabella Thomas)[8] (THOMAS et al. 2012) répondait aux besoins d'amélioration de la qualité de l'information et de sécurité des usagers et des

[7] ANR-06-SECU-007, sous la direction de Sylviane Cardey, en collaboration avec Airbus France et le Centre Hospitalier Universitaire de Besançon. Le projet LiSe répondait aux plusieurs objectifs distincts, concernant notamment la traduction automatique et le 'data mining' que je ne mentionne pas ici. Pour voir la bibliographie complète du projet : http://projet-lise.univ-fcomte.fr/publications.html.

[8] Site du projet : http://tesniere.univ-fcomte.fr/sensunique.html

personnels, quelle que soit la branche d'activité. Il s'appuyait sur le concept de Langue Contrôlée (LC) développé dans le projet LiSe afin d'améliorer la qualité des textes techniques et d'en optimiser la rédaction (Figure 1). Les objectifs du projet Sensunique ont été organisés en trois volets :

- optimiser la méthodologie de contrôle de la langue et valider l'opérationnalité de la rédaction en LC sur un ensemble de textes issus du domaine de la santé ; mettre en place des procédés d'aide pour accélérer l'établissement du lexique d'une LC ;
- optimiser le logiciel d'aide à la conception et à la rédaction en LC de textes techniques (appelé Rédacticiel Prolipsia, du nom de la start-up qui l'a développé) afin d'améliorer la réponse apportée aux besoins des rédacteurs, mais aussi de le rendre plus rapidement déclinable à divers domaines et types de textes ;
- valoriser et populariser le concept de LC et tester l'acceptabilité du rédacticiel à travers des ateliers de sensibilisation et d'échanges, auprès de divers utilisateurs et prescripteurs potentiels (professionnels de Santé et autres).

Pour consolider le processus de rédaction de textes techniques en LC, nous avons développé un logiciel web d'aide à la conception et rédaction en LC (le Rédacticiel Prolipsia). Conçu pour guider la rédaction initiale de protocoles, il se compose :

- d'un gestionnaire de LC, destiné aux linguistes qui vont recenser et formaliser en base de données les unités du lexique autorisées ou interdites (lexies spécialisées ou non), les structures linguistiques (phrastiques et lexicales), définir les architectures de texte et les règles de rédaction ;
- d'une interface de rédaction, pensée pour assister le rédacteur pas à pas, sans connaissances linguistiques préalables. Le rédacteur a aussi la possibilité d'ajouter des unités lexicales, à condition que celles-ci ne comportent aucun risque d'ambigüité (par exemple : noms de marque, noms de produits chimiques, toponymes, etc.).

Dans le cadre du projet Sensunique, nous avons développé des modules logiciels (décrits plus loin) afin d'optimiser le Rédacticiel.

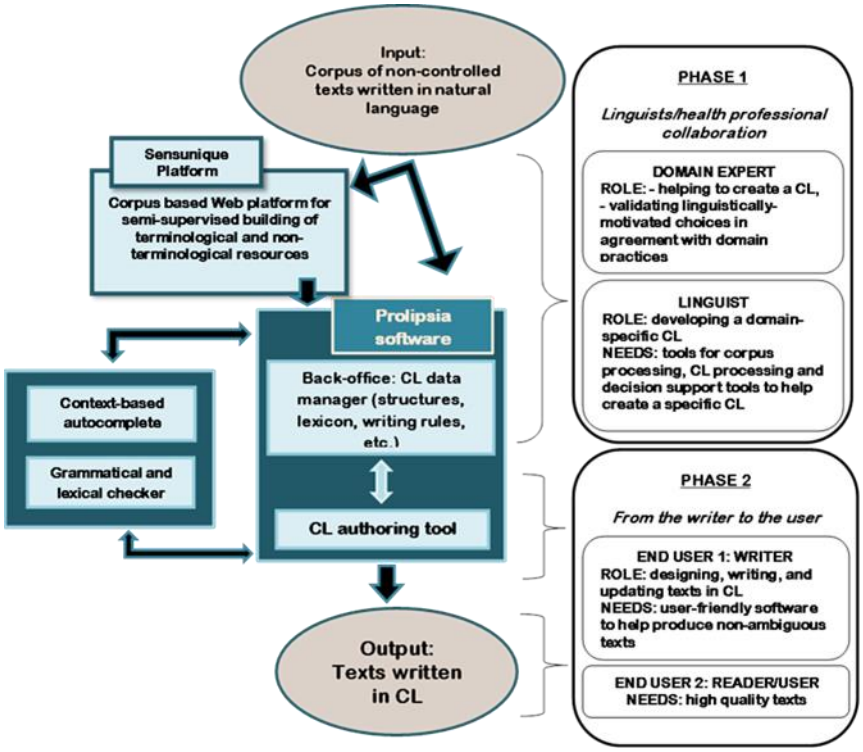


Figure 1 Schéma du projet Sensunique

Pour optimiser le procédé d'écriture en LC assistée par un rédacticiel, nous avons mis en place un travail interdisciplinaire autour de trois types de compétences :

- linguistiques et TAL (Centre L. Tesnière) ;
- informatiques (FEMTO-ST/ DISC) ;
- compétences métier, liées, dans le projet, au domaine de la Santé.

Pour bénéficier de façon optimale des compétences professionnelles en Santé, le partenariat académique avec la SFR FED 4234 a été complété par des accords avec l'EFS B/FC) et le CHRU de Besançon, des structures interrégionales majeures dans les secteurs du diagnostic et des soins, en lien avec les instances nationales. En collaboration avec ces institutions, et pour optimiser la méthodologie d'établissement de LC, nous avons décidé de travailler sur deux domaines d'applications :

- Immunobiologie (Modes Opératoires en immunobiologie), avec l'appui de l'EFS B/FC ;
- Médecine Nucléaire (Protocoles de gestion des déchets), avec l'appui du CHRU de Besançon.

Pour la valorisation de la LC, nous avons organisé, en collaboration avec Vuitton Consultant, la société Prolipsia et l'Institut Édouard Belin, neuf ateliers d'échanges entre professionnels, en France et en Suisse francophone. Chacun de ces ateliers était destiné à un public homogène, soit par le métier (qualiticiens, biologistes, gestionnaires du risque…), soit par les préoccupations (communication en situations d'urgence ; accréditation des laboratoires de biologie médicale, recherche clinique…).

Nous avons mené parallèlement les missions suivantes :

- conception et développement des logiciels, afin d'optimiser le procédé de contrôle de la langue, aussi bien pour les besoins de l'ingénieur linguiste que du rédacteur technique ;
- travail de terrain, sur un corpus et avec des professionnels de Santé, servant de base à l'amélioration de la méthodologie d'établissement des LC ;
- création d'un réseau d'utilisateurs et de prescripteurs grâce aux ateliers d'échanges, mentionnés ci-dessus.

Nous avons conçu et développé deux logiciels (déposés à l'APP[9]) durant le projet :

Station Sensunique :

IL s'agit d'une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de vocabulaire (orientée LC). Elle a pour objectif d'accélérer le processus d'établissement du lexique d'une LC, en s'appuyant sur la collaboration de plusieurs outils TAL et en interrogeant automatiquement des ressources terminologiques existantes ; elle permet aussi la gestion du lexique et l'export des dictionnaires. La Station Sensunique est mise à disposition gratuitement (http://www.station-sensunique.fr/) pour une utilisation à des fins de recherche et/ou d'enseignement.

Modules Sensunique :

Deux modules ont été développés en vue de leur intégration au rédacticiel Prolipsia, pour faciliter d'une part, le travail d'un linguiste (le module de Gestionnaire de Données Sensunique, GDS) et, d'autre part, celui du rédacteur (le module d'Optimisation de la Saisie du Rédacteur, OSR).

Les travaux entrepris avec l'EFS B/FC et le CHRU de Besançon sur les LC ont permis :

- la mise en place d'une méthodologie de travail entre experts métier et linguistes ; cette méthodologie, établie et validée dans le domaine Santé, est transposable à tous les domaines d'application ;
- la rédaction en LC de protocoles, modes opératoires et instructions représentatifs de l'activité d'immunobiologie de l'EFS B/FC et de celle de la gestion des déchets radioactifs au CHRU de Besançon ;
- une prise de conscience des professionnels de la Santé quant à la nécessité d'améliorer la qualité des textes et de standardiser leur écriture. C'est un préalable important à la diffusion commerciale du service linguistique et du rédacticiel proposés.

Le projet a aussi permis de constituer un réseau de plus de 50 institutions/ entreprises utilisatrices potentielles, en Santé, mais aussi agro-alimentaire, banque/assurance, ingénierie, administration, ainsi que d'établir ou renforcer plusieurs collaborations à l'international.

[9] Agence de Protection de Programmes.

2.2 Écrits scientifiques

2.2.1 Projet SARS

Le projet SARS, *Système d'aide à la rédaction scientifique (SARS) dans le domaine biomédical* (2015-2017, en cours), que je coordonne, est soutenu par le Conseil Régional de la Franche-Comté, en collaboration avec FED 4234 (responsable scientifique : Professeur Estelle SEILLES). Pour répondre aux besoins précis de chercheurs francophones devant rédiger des textes scientifiques en anglais, ce projet vise à développer une aide informatisée à la rédaction des articles, sous forme d'une interface interactive d'édition. En consultation avec notre partenaire santé, nous avons choisi le thème des essais cliniques pour conduire les premières expérimentations sur la possibilité d'une mise en place d'un tel outil. Nous nous appuyons sur des corpus de textes existants (un corpus de 100 000 articles publiés dans PubMed et PLOS en accès libre) et sur un corpus sur les essais cliniques construit expressément pour le projet pour en extraire des expressions structurant le discours et l'argumentation scientifique, des reformulations, des exemples d'utilisation en corpus ainsi que des lexiques terminologiques liés au domaine.

Le projet s'articule autour des tâches suivantes :

- Réflexion sur les pratiques et les besoins des utilisateurs, avec les enseignants-chercheurs de la FED 4234, traducteurs employés pour aide à la traduction par le CHRU de Besançon, experts en rédaction médicale (intervenants et membres du conseil scientifique du projet : Hervé MAISONNEUVE[10], Jean-Lou JUSTINE[11]) ; identification des corpus d'analyse appropriés ; établissement d'un cahier des charges ;
- Analyse automatique des corpus scientifiques dans le domaine biomédical afin d'identifier la structure rhétorique des différentes sections, les catégories sémantiques présentes dans les textes et les formes de surface (expressions linguistiques) correspondantes ; modélisation linguistique des informations obtenues en vue de création d'une interface de rédaction assistée ;
- Acquisition de ressources terminologiques par la méthode de multi-extraction, développée au Centre Tesnière (ANR Sensunique) et implémentée initialement pour la langue française dans la Station Sensunique. L'extension de cette plateforme vers la langue anglaise permettra d'acquérir des bases lexicales et vocabulaires terminologiques directement à partir de textes rédigés en anglais ;
- Validation et évaluation des résultats.

Le projet est toujours en cours, mais certains de ces résultats sont décrits dans la section 4.1.

[10] Engagé dans l'enseignement de la rédaction médicale depuis 1975, rédacteur en chef-adjoint de La Presse Médicale, membre depuis 2002 du 'Rome CME/CPD group', auteur du livre le plus consulté par les rédacteurs francophones : 'La Rédaction Médicale' ; http://www.h2mw.eu/

[11] Rédacteur en chef du journal en ligne en langue anglaise 'Parasite', http://www.parasite-journal.org/

2.2.2 Plateforme d'apprentissage du lexique spécialisé

Le projet d'une plateforme pour l'apprentissage du lexique spécialisé a débuté avec le mémoire de recherche de Master 2 de François-Claude REY (soutenu en juin 2016), que j'ai encadré. Il s'agit de créer un support logiciel pour les enseignants de langues étrangères de spécialité de niveau académique. Le projet vise à répondre à un besoin détecté lors de la préparation de cours de langues étrangères de spécialité. Il s'agit de doter les enseignants de langues de moyens automatisés et semi-automatisés pour préparer rapidement et avec le moins de tâches répétitives possible les supports de cours : textes, listes de vocabulaire et exercices. La plateforme doit générer automatiquement des exercices d'apprentissage de langues spécialisées, à partir d'un texte de langue étrangère de spécialité fourni par l'enseignant et à partir des connaissances du domaine intégrées dans la plateforme. Pour la mise en place d'un prototype, un type d'exercices a été choisi— les exercices à trous— ainsi qu'un domaine de langue de spécialité de niveau académique — l'anglais de la géographie de l'eau.

Le projet est toujours en cours, mais certains de ces résultats sont décrits dans la section 4.2.

3. APPROCHES NORMATIVES EN REDACTION TECHNIQUE ET CONCEPT DES LANGUES CONTROLEES

La volonté grandissante de normalisation, standardisation et les besoins de réglementation (entre autres choses, pour des raisons juridiques ou de sécurité) fait augmenter considérablement le nombre de textes procéduraux dans tous les domaines professionnels. Ces textes soulignent l'obligation d'accomplir des actions précises (de respecter des consignes précises) dans des contextes très cadrés.

La qualité et l'exactitude de l'information et l'efficacité de la transmission peuvent s'avérer vitales quand il s'agit de santé, de sécurité ou dans tout autre domaine, que l'on soit en situation de crise ou non. Or, de trop nombreuses erreurs d'interprétation des textes "mal" écrits mènent à des actions inappropriées, parfois lourdes de conséquences (CARDEY 2009; VUITTON et al. 2009 ; RENAHY et al. 2011 ; THOMAS et al. 2015). Il devient alors primordial d'améliorer la qualité des textes procéduraux produits en amont afin de minimiser en aval le risque d'erreurs, tout en réduisant le temps d'appropriation de la procédure.

Rédiger selon les principes de la Langue Contrôlée (LC) permet d'obtenir des textes techniques clairs, structurés et exempts d'ambigüité. Les avantages sont nombreux : parce que le message est univoque, la compréhension est facilitée, y compris en cas de stress. Les textes obtenus sont uniformisés, donc réutilisables. Ils sont conformes à des modèles et normes établis, et répondent à des critères de qualité permettant de plus leur traitement par des outils de TAL (notamment dans l'optique de faciliter la traduction, humaine ou automatique).

Cependant, malgré les nombreux avantages de rédaction en LC, l'établissement, l'apprentissage et l'utilisation effective des LC ne va pas de soi pour les rédacteurs, qu'ils soient professionnels ou non. Plusieurs écueils empêchent la propagation de cette technique parmi les rédacteurs, entre autres :

- la quantité de travail nécessaire pour établir une LC ;
- le temps d'apprentissage des règles d'une LC ;
- l'effort rédactionnel pour répondre aux exigences d'une LC (par exemple, utiliser seulement les structures et le lexique autorisés, utiliser un style répétitif et contraint etc.).

Une des façons pour pallier ces difficultés consiste à proposer aux rédacteurs des outils automatisés d'aide à l'écriture en LC. C'est pourquoi nous avons conçu une interface logicielle d'aide à la rédaction, *Compagnon LiSe* (RENAHY et al. 2009) avec pour objectif de faciliter la rédaction des textes procéduraux par des rédacteurs techniques non familiers avec les exigences d'une LC. Dans la suite de ce chapitre, nous allons décrire la méthodologie qui nous a amenés à proposer les principes du fonctionnement de cet outil, conçu lors du projet LiSe, ainsi que les développements apportés par le projet Sensunique. Au cours du projet Sensunique nous nous sommes particulièrement intéressés au problème de l'établissement du lexique d'une LC : nous avons conçu et développé une plateforme d'aide au recensement de ce lexique, appelée *Station Sensunique* que nous allons aussi détailler dans la suite de ce

chapitre. Mais commençons d'abord par un bref état de l'art sur les LC et surtout sur les outils d'aide à leur écriture.

3.1 État de l'art sur les Langues Contrôlées et outils d'aide à leur écriture

Le langage est un outil incontournable pour la transmission d'informations précises, non ambigües, susceptibles d'être partagées par un grand nombre d'intervenants, capables de se traduire en actions immédiates, rapides et coordonnées. Le Conseil Européen, dans le document « Accord partiel ouvert en matière de prévention, de protection et d'organisation des secours contre les risques naturels et technologiques majeurs »[12] s'est attaché à rédiger un glossaire des mots et des expressions en relation avec les risques majeurs. Malgré un titre évocateur « La gestion linguistique du risque », il ne s'agit pourtant que d'un dictionnaire qui donne les équivalents terminologiques français, anglais et allemand, mais en aucun cas des recommandations pour la rédaction des supports de communication en cas de crise. La linguistique y est clairement un aspect très négligé. On évoque toujours les problèmes de communication, mais le support linguistique de cette communication, pourtant essentiel, n'est jamais signalé. Il est, pour le moins, surprenant que les aspects linguistiques ne soient pas pris en compte alors que de nombreuses incompréhensions, et donc des réactions inappropriées, sont souvent liées à des problèmes linguistiques assez triviaux qui auraient pu être aisément évités par une préparation adéquate.

Pourtant, le besoin de simplification et de standardisation n'est pas nouveau. Dès début du XX siècle des formules de « lisibilité » ont été mises en place, principalement dans le but de rendre plus accessibles les manuels scolaires utilisés aux Etats-Unis (DUBAY 2004). Bien qu'il ne s'agisse pas là de LC, on peut y voir un début, du moins une prise de conscience des difficultés liées aux langues naturelles.

Les LC ont été d'abord développées pour répondre aux problèmes de compréhensibilité dans la communication orale et écrite en milieu multilingue et réduire les coûts imputables à la traduction de modes d'emploi et manuels de maintenance. La plupart des exemples de LC, qu'elles soient « orientées machine » (c'est-à-dire dans l'optique d'un traitement automatique du contenu) ou « orientées humain » (dans le but d'améliorer la lisibilité et la compréhensibilité), provient essentiellement de l'industrie privée où ces langues ont été développées habituellement « en interne », pour répondre à des besoins propres à ces entreprises. De ce fait, les données relatives aux LC développées sont relativement confidentielles, et l'information accessible permet d'appréhender les LC dans les grandes lignes uniquement (WU 2005 ; GAVIEIRO-VILLATTE et al. 1999 ; BARTHE 2004).

La recherche scientifique relative aux LC est largement dominée par l'anglais. Kuhn (2013) recense plus de 100 LC conçues dans cette langue. Ceci semble évident, étant donnée la présence internationale de cette langue, en particulier dans les domaines techniques et notamment dans l'aéronautique, domaine par essence multilingue, précurseur en la matière. Toutefois, on utilise aujourd'hui quelques LC basées sur d'autres langues que l'anglais et appliquées à d'autres domaines que l'aéronautique, comme par exemple une LC suédoise

[12] Accord européen et méditerranéen sur les risques majeurs, http://www.coe.int/fr/web/europarisks [29/09/2016].

Scania (ALMQVIST 1996). Très peu d'études ont été menées sur le contrôle de la langue française, et bien qu'il existe un Français Rationalisé, développé pour l'industrie aéronautique, celui-ci a été établi à l'origine d'après le guide de rédaction du Simplified English, c'est-à-dire d'après des règles de rédaction établies spécifiquement pour l'anglais. Ceci peut laisser dubitatif quant au résultat, et particulièrement si on connaît les divergences linguistiques qui existent entre ces deux langues, ce qui a d'ailleurs retardé considérablement l'aboutissement du projet (LESEIGNEUR 1999). Le projet LiSe s'est ainsi démarqué des autres projets de recherche et a significativement contribué à l'augmentation de l'état de l'art des LC en développant une méthodologie de contrôle de la langue basée sur le français et applicable à différents domaines (LC LiSe).

La plupart des grands groupes industriels ayant produit leurs propres LC ont également développé leurs propres analyseurs ('LC checkers'). Nous pouvons citer :

- les analyseurs pour PACE et pour ScaniaSwedish, de Perkins Engine Ltd. ;
- l'analyseur SECC (Simplified English Grammar and Style Checker/Corrector) pour la LC d'Alcaltel Telecom, né du Projet CEC-funded SECC ;
- les analyseurs pour l'industrie aéronautique : BSEC (Boeing Simplified English Checker) pour la LC STE (Simplified Technical English), EGSC (Enhanced Grammar, Style and Content Checker), l'analyseur Eurocastle pour la LC SE (Simplified English) conçu dans le projet GRAAL ; AlethCL, LANTmaster, ClearCheck, MAXit.

Même si en France, le concept d'outil d'aide à la rédaction de textes techniques en LC est complètement nouveau, au niveau européen, on trouve déjà des outils de ce type, et principalement deux correcteurs de LC :

- Acrocheck, développé par la société allemande Acrolinx pour les LC suivantes : Simplified Technical English, Standard Technical German, Standard Technical Japanese, Global English, English for non-native speakers ;

- HyperSTe : développé par la société hollandaise Tedopres[13] qui s'adresse aux industries utilisant la LC de l'AECMA (Simplified English).

En France, le projet ANR Lelie[14] a donné naissance à un logiciel d'assistance à la rédaction technique, qui a pour objectif « la détection de sources d'incompréhensions potentielles liées à la qualité de la langue dans des textes procéduraux » (BARCELLINI et al. 2014). Ce logiciel est implémenté dans une plateforme TEXTCOOP (SAINT-DIZIER 2012) qui est un environnement basé sur les grammaires logiques, dédié à l'analyse de structures discursives. Il analyse la qualité de la langue dans les documents procéduraux à trois niveaux :

- linguistique (expressions complexes, informations implicites etc.) ;
- incohérence métier (façons inhabituelles de réaliser une action) ;

[13] Devenue Suédoise après son rachat par Etteplan.
[14] ANR-10-EMMA-0011, *Un logiciel intelligent d'aide au diagnostic de risques dans les procédures industrielles* ; coordinateur du projet : Patrick Saint-Dizier.
.

- non-respect d'exigences métier et de sécurité (en comparaison avec les consignes ou avertissement de sécurité indiqués dans une base d'exigences métier).

Selon un rapport de l'UE sur le marché européen de l'industrie de la langue (RINSCHE 2009), Acrolinx estime le marché pour ce genre d'outils entre 7 et 10 millions d'Euros, sans compter sur l'émergence de nouveaux marchés, tel que celui de la santé et sur la prise de conscience de l'impact majeur que peut avoir la linguistique sur la qualité et la sécurité globale. Il est vrai que parmi les répondants à l'étude menée par l'UE, seulement la moitié avait connaissance de l'existence d'outils d'aide à la rédaction en LC, et seulement 1,6% en utilisaient un, témoignant de la méconnaissance générale sur le sujet.

3.2 Paramétrage d'une Langue Contrôlée

Les recherches sur les LC montrent qu'il ne peut exister de LC universelle[15]. Les différentes LC qui ont été développées pour l'anglais ne sont pas identiques, puisque leur conception dépend de différents paramètres. En plus de contraintes typiques aux LC, telles que la clarté, la lisibilité et la non-ambigüité, la création d'une LC prend en compte les critères liés aux types de documents à rédiger, aux domaines concernés, au public cible et à la traduction automatique (O'BRIAN 2003). Chacun de ces critères impose des contraintes que nous avons synthétisées pour la LC LiSe dans le schéma suivant (Figure 2) :

[15] Ce chapitre est basé sur RENAHY et al. (2009).

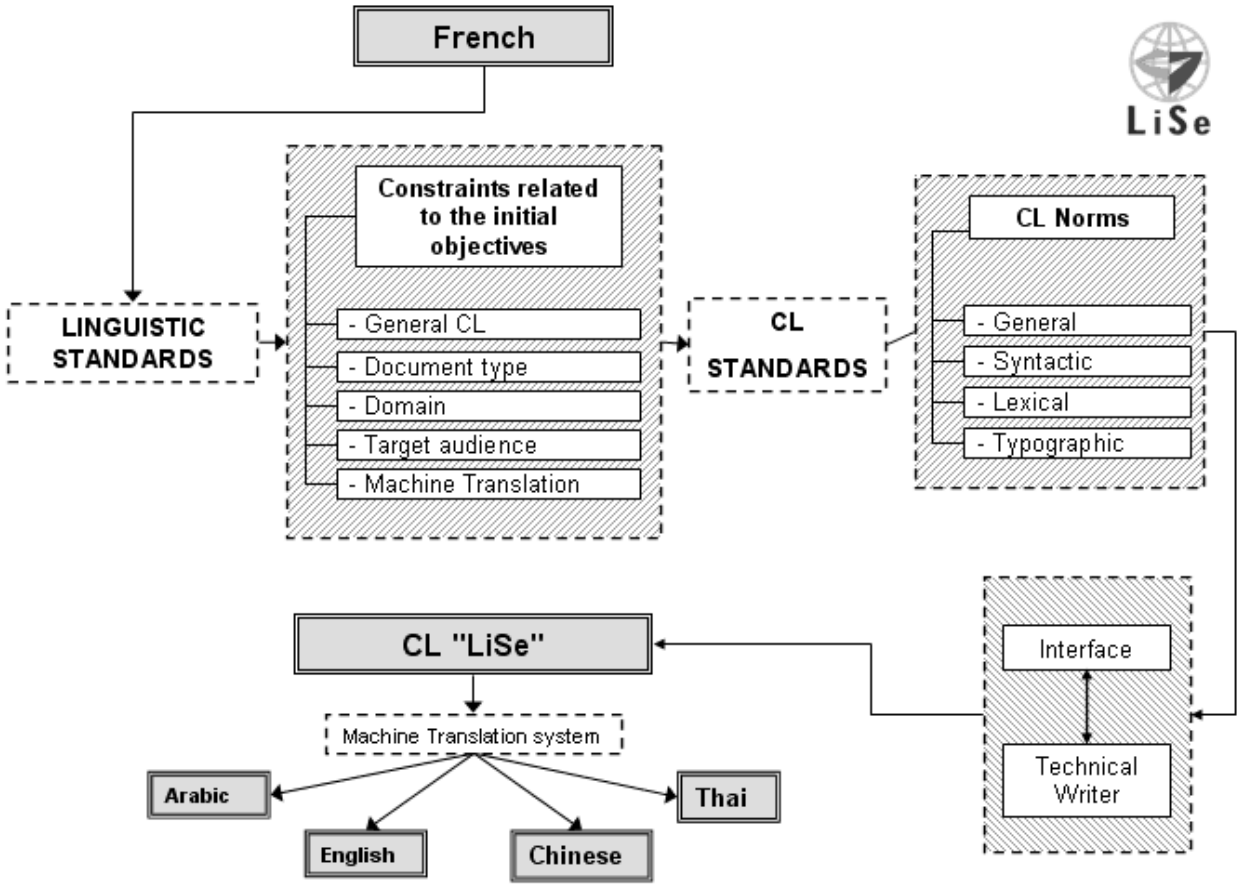


Figure 2 Paramètres d'établissement de LC LiSe

Pour établir la LC LiSe, chaque norme linguistique a été considérée sous l'angle de différents critères (critères généraux et critères spécifiques : type de document, domaine, public cible, traduction automatique). Ceci a mené à l'établissement de l'ensemble des règles de contrôle, lesquelles à leur tour ont été subdivisées en quatre catégories : règles générales, typographiques, syntaxiques et lexicales.

Les **critères généraux** liés à l'établissement d'une LC (clarté, lisibilité, non-ambiguïté) imposent aux rédacteurs des restrictions sur la façon d'écrire les documents qui concernent :

a) la structuration de l'information ;

Exemples de règles :

L'information doit respecter l'ordre logique et chronologique.

Une phrase ne peut contenir qu'une seule information ou action à effectuer.

Même si certaines règles semblent évidentes et relèvent du bon sens l'analyse des corpus que nous avons menée montre qu'il est toujours utile de les rappeler aux rédacteurs, puisqu'on trouve dans les textes des erreurs provenant de non-respect de ces règles (RENAHY et al. 2011 ; THOMAS et al. 2015).

b) le contrôle syntaxique ;

Exemples de règles :

Une phrase ne peut contenir qu'un seul verbe.

Les structures verbales autorisées sont prédéfinies en ce qui concerne le nombre et la nature de leurs compléments.

Ces types de règles peuvent sembler très restrictifs, mais ils s'avèrent appropriés dans le contexte de rédaction de textes procéduraux. En plus, ils facilitent le respect des autres règles, plus 'floues', telles que 'une phrase = une information ' ou 'un mot = un sens'. Effectivement, une définition précise d'une structure verbale permet d'en délimiter le sens, comme dans l'exemple suivant :

Qqch passer sous Qqch (« si la fumée passe sous la porte »)

v/s

Qqn passer à Qqch (« passer à l'étape suivante »)

c) le contrôle lexical ;

Exemple de règles :

Une unité lexicale ne peut avoir qu'un seul sens.

Cette règle vise à éliminer l'ambigüité lexicale. Lorsqu'un mot est polysémique, il doit être utilisé avec un contexte désambigüisant, comme par exemple, le mot *prise* en français **:** *prise électrique, prise de judo, prise de médicaments*[16]. Si le contexte désambigüisant n'est pas disponible, un seul sens est choisi et autorisé par unité lexicale.

Les **contraintes spécifiques** liées aux types de documents peuvent varier en fonction de ceux-ci. Le projet LiSe a pris en charge un type particulier de documents, à savoir, les documents de type protocolaire. Nous les avons définis de façon suivante (définition basée sur HEURLEY 2001/2 et BOUFFIER 2006) :

Un protocole est un texte qui vise à communiquer à un utilisateur, qu'il soit ou non spécialiste du domaine, les actions à accomplir pour attendre un objectif, sous certaines conditions. Un protocole peut être destiné à une exécution immédiate, avec des différents niveaux d'urgence, ou il peut instruire un utilisateur potentiel sur les actions à entreprendre en cas d'urgence.

Il est important de savoir que la personne à qui est destiné un protocole, n'est pas simplement un 'lecteur' de celui-ci mais surtout un 'exécuteur'. L'objectif du protocole est de faire faire à la personne les actions précises, décrites dans le protocole, d'où l'importance de la manière dont les informations sont transmises. De plus, même si la plupart des protocoles sont écrits pour être consultés avant qu'un évènement n'arrive, la majorité des utilisateurs ne s'y réfèrent pas avant d'en avoir réellement besoin, c'est-à-dire souvent dans un contexte d'urgence et de stress (lors d'un incendie ou d'un accident par exemple).

[16] Nous reviendrons sur la question du lexique dans le &3.4.

Nous avons identifié et formalisé la structure interne des protocoles en nous appuyant sur un corpus de 450 protocoles appartenant à deux domaines d'application : la sécurité civile et le biomédical (ceux-ci fournis par le CRHU de Besançon). Nous avons mis en place la notion de structures fonctionnelles, c'est-à-dire des structures ayant une fonction spécifique dans un texte protocolaire. Chaque structure fonctionnelle active un ensemble fini de structures syntaxiques et de règles de mise en page[17]. Les cinq principales structures fonctionnelles sont:

- ***Titre*** : une proposition courte donnant le titre du protocole ; sa construction est restreinte à 5 structures syntaxiques ;

Exemple :
Structure : 'Que faire en cas de' + Gn_sa + '?'
Réalisation : Que faire en cas d'incendie ?

- ***Sous-titre*** : une proposition courte qui introduit les sections ou les sous-sections spécifiques d'un protocole ; les mêmes restrictions syntaxiques que pour le titre, mais une mise en forme différente ;

Exemple :
Structure : Gn_sa.
Réalisation : Ouragan.

- ***Instruction*** : une phrase injonctive décrivant une action qui doit être exécutée par l'utilisateur. Nous avons normalisé les instructions en imposant l'utilisation d'un verbe à l'infinitif pour décrire une injonction. Par conséquent, toutes les instructions commencent par un verbe à l'infinitif.

Exemple :
Structure : opt(neg(Neg)) + Vinf + Arg1+ '.'
Réalisation : Nettoyer les conduits.

- ***Condition*** : une phrase conditionnelle rependant au schéma « Si X alors Y » avec ses variantes, par exemple conditions coordonnées (« Si X et Y alors Z ») et conditions disjointes (« Si X ou Y, alors Z ») ; les conditions permettent l'utilisation des verbes conjugués.

Exemple :
Structure : 'Si' + Arg0 + Vconj + ' :'
Réalisation : Si l'ouragan approche :

- ***Note explicative*** : cette structure fonctionnelle permet au rédacteur d'expliquer les raisons de certaines instructions, s'il le juge nécessaire pour amener l'utilisateur à effectuer l'action. Il s'agit typiquement de justifier un geste dont les motifs ne seraient pas évidents pour l'utilisateur ou un geste qui serait contraire aux habitudes. Les notes explicatives sont toujours précédées par un mot introducteur qui en définit le contenu ('Exemple', 'Objectif', 'Explication' etc.), mais leur utilisation est limitée pour éviter de surcharger les textes avec des informations superflues, qui peuvent distraire les utilisateurs.

[17] Par contre, les règles lexicales ne sont pas dépendantes de structures fonctionnelles.

Exemple :
Structure : ‘Explication’ + ‘ : ’+ Gn_sa + '.'
Réalisation : Explication : Risque d’étouffement.

À part les structures fonctionnelles prédéfinies pour la LC LiSe, aucune autre n’est permise lors de l’écriture d’un protocole. Ceci assure que l’information est décrite correctement et de manière homogène. De plus, une mise en page spécifique associée à chaque structure permet de distinguer rapidement entre les différents types d’informations et d’améliorer ainsi la lisibilité du document.

Les **contraintes dépendantes du domaine** sont divisées en trois types. D’abord les contraintes liées **à la sécurité** : plus le domaine relève de la haute sécurité, plus il exige des instructions précises et une compréhension immédiate, plus le contrôle devient strict. La LC LiSe est un exemple d’une LC avec de très strictes règles de contrôle. Le second type de contraintes est lié aux **normes de rédaction** établies dans un domaine. Les règles d’une LC doivent respecter autant que possible les pratiques prévalant dans un domaine, pour ne pas paraitre artificielles et difficiles à accepter par les rédacteurs et les utilisateurs. Par conséquent, il est nécessaire de les faire valider par un spécialiste du domaine. Le troisième type de contraintes concerne les **contraintes linguistiques** liées au domaine, et notamment le lexique qui diffère naturellement d’un domaine à l’autre. Par conséquent, la définition d’une unité lexicale n’est valable qu’à l’intérieur d’un domaine et une même unité lexicale peut avoir une définition différente dans deux domaines distincts.

Exemple
Sécurité civile : raviver$_1$ (‘rendre un feu plus vif’) :
Ne pas raviver$_1$ les braises d’un barbecue avec de l’alcool.
Biomédical : raviver$_2$ (‘remettre à nu la chair vive’) :
Raviver$_2$ la plaie.

Plus surprenant, même si les structures syntaxiques autorisées sont largement partagées à travers les différents domaines, il est néanmoins nécessaire d’inventorier des structures très spécifiques, par exemple pour le domaine médical, dans lequel le nombre et l’ordre de chaque complément du nom est important et peut être prédéterminé.

Exemple
1 seringue de 5 mL d'Alteplase 2 mg / 2 mL.
2 ampoules de 10 mL de NaCl injectable 0,9 %.

Le dernier type de contraintes, contraintes liées à la **traductibilité**, est imposé pour assurer une bonne qualité de traduction automatique des protocoles rédigés en LC LiSe. Pour répondre au besoin d’économie, d’ergonomie et d’urgence (pour les domaines à haute sécurité), il a été décidé de créer une langue contrôlée unique pouvant être traduite en plusieurs autres langues, aussi différentes que l’anglais, le thaï, l’arabe et le chinois. Par conséquent, chaque règle de contrôle a été examinée en tenant compte de son impact sur la traduction en chacune de ces quatre langues. Ceci a résulté en trois types de contraintes[18] :

[18] Les résultats détaillés pour l’arabe et le chinois sont décrits respectivement dans les thèses de M. BEDDARD (2013) et de G. JIN (2015) ; voir aussi les publications du projet LiSe (http://projet-lise.univ-fcomte.fr/publications.html).

a) **contraintes syntaxiques** : l'objectif étant de réduire au maximum les divergences syntaxiques entre la langue source et les langues cibles, certaines constructions ont été interdites en français à cause de la difficulté de leur traduction (par exemple, le passif, jugé trop compliqué à traduire automatiquement en arabe).
b) **contraintes lexicales** : les unités lexicales ont été passées au crible de leurs possibles traductions et désambiguïsées ou interdites en fonction de leurs équivalents en langues cibles.

> Exemple
> Même si, en français, il est possible d'utiliser la construction: jeter + liquide (jeter de l'huile, de l'eau), cette construction a été interdite à cause du chinois et du thaï, qui font la distinction nette entre jeter + solide et verser + liquide.

c) **contraintes morphologiques** : les contraintes morphologiques limitent l'utilisation de certains temps et modes verbaux, jugé non nécessaires en français (pour la rédaction de protocoles) et compliqués à traduire.

Il est important de savoir qu'à chaque fois qu'une contrainte est établie, une solution de remplacement est proposée au rédacteur, pour éviter les situations de blocage de l'écriture.

L'ensemble de règles de contrôle établies pour la LC LiSe a été rassemblé dans un manuel utilisateur (lui-même rédigé en langue contrôlée). Plus de 150 règles de contrôle sont consignées dans ce manuel et ce nombre n'inclut pas les règles de contrôle lexical. Toutes les règles sont nécessaires pour assurer le bon fonctionnement de la LC LiSe, mais leur nombre peut être décourageant pour les rédacteurs potentiels : il n'y a aucune utilité à construire une LC avec des règles si volumineuses qu'il est impossible de se les rappeler et de les appliquer.

L'idée d'informatiser le processus de l'écriture en LC provient de ce constat et du questionnement qui le suit : comment faciliter la tâche de rédaction en LC à un rédacteur, qui ne connait pas le concept de la LC et qui, de plus, n'est pas un rédacteur technique formé ? Nous décrivons notre solution dans le chapitre suivant.

3.3 Informatisation du processus d'écriture en Langue Contrôlée : génération de textes en LC

Les langues contrôlées sont considérées par les rédacteurs techniques comme difficiles à apprendre et à utiliser. Même si les textes écrits en LC gagnent en clarté et en lisibilité par rapport à des documents plus 'classiques'[19], le nombre de restrictions qu'elles imposent est un obstacle considérable à leur propagation. On observe une réticence par rapport à l'utilisation des LC de la part des rédacteurs (ALLEN 2003 ; CARDEY 2009), bien souvent professionnels de leur domaine, mais sans aucune formation en rédaction technique. Le maniement d'une LC présuppose non seulement l'apprentissage des règles de contrôle, mais aussi et avant tout le changement des habitudes rédactionnelles, ce qui constitue un obstacle majeur à l'adoption d'une LC. Les rédacteurs peuvent se sentir limités voire bloqués par l'utilisation d'une LC tant ils sont habitués à un certain style d'écriture, inculqué depuis l'école,

[19] Voir l'exemple d'un document contrôlé lors du projet Sensunique en Annexe 9.1.

où l'on nous apprend à « *ne pas répéter les mots* », à « *ne pas utiliser de phrases nécessairement courtes* », à accorder plus d'attention à « *faire élégant et joli* » (KLINKENBERG 2002). Cette attitude très « française », mais que l'on peut retrouver sous d'autres formes dans d'autres langues, a été un frein important à la diffusion des connaissances issues de la recherche française dans la littérature scientifique internationale.

Le logiciel d'aide à la rédaction que nous avons conçu durant le projet LiSe – appelé le Compagnon LiSe - devait tout d'abord faciliter l'approche, l'apprentissage et l'utilisation de la LC LiSe, particulièrement par un rédacteur qui n'était pas un spécialiste en linguistique. Pour ce faire, cet outil devait prendre en charge les règles de l'écriture des protocoles en LC à travers une interface facile à accepter par les utilisateurs potentiels, c'est-à-dire intuitive, conviviale, facile à manipuler, et n'exigeant pas de connaissances de notions linguistiques avancées. Il devait accompagner le rédacteur tout au long du processus de rédaction, tout en assurant, dès l'étape même de la rédaction jusqu'à la mise en forme du texte final, la conformité des textes obtenus avec les règles de contrôle. En effet, contraindre dès le départ l'entrée du texte évite au rédacteur une relecture et correction qui peuvent être pénibles, même si elles sont assistées par ordinateur. Une troisième fonction, en plus de l'aide à la rédaction et de la vérification de conformité, a été assignée à ce logiciel. Il s'agit de la structuration des textes par le rédacteur en vue de leur traduction automatique. Le logiciel amène le rédacteur, en s'appuyant sur la structure de la langue (grammaticale, syntaxique, lexicale) à effectuer lui-même un certain nombre d'opérations mises en mémoire par la machine et indispensables à produire une bonne traduction.

De l'autre côté, comme pour tout logiciel traitant de la langue naturelle, l'information encodée doit être traitée et formalisée, ce qui limite fortement les choix de l'utilisateur lors de la rédaction de protocoles, mais est indispensable pour assurer la qualité de la rédaction et de la traduction.

Pour résumer, le Compagnon LiSe, tout en restant ergonomique, doit prendre en charge des exigences de systèmes informatisés et rester dans les limites acceptables par l'utilisateur pour répondre aux objectifs suivants :

- faciliter l'acquisition et la manipulation de la LC LiSe ;
- assurer la conformité des protocoles rédigés avec les règles de la LC LiSe, c'est-à-dire aider à obtenir des textes de qualité ;
- préparer les données pour la traduction automatique en 'collaborant' avec le rédacteur technique.

Nous décrivons l'interface que nous avons conçue dans la section qui suit.

3.3.1 Compagnon LiSe : une interface collaborative d'aide à la rédaction en LC

Comme suggéré précédemment, une interface qui doit faire face aux exigences d'une LC, d'un système de traduction automatique et être à la portée de n'importe quel rédacteur technique n'est pas une chose facile à mettre en place. De fait, le Compagnon LiSe n'est pas

un simple système d'aide à la rédaction, mais aussi un outil collaboratif, qui implique l'utilisateur dans la préparation des données pour la traduction automatique[20].

L'interface dispose de deux profils utilisateurs, désignés en fonction de compétences et de l'habilité d'un utilisateur potentiel, qui, rappelons-le :

- n'est pas forcément un rédacteur technique professionnel ;
- n'est pas forcement familier avec le concept de la LC ;
- ne possède pas forcement des connaissances poussées en linguistique ;
- pourrait être amené à rédiger en urgence.

Le profil **utilisateur novice** permet à l'utilisateur d'être guidé pas à pas tout au long du processus de la rédaction, alors que le profil **utilisateur expert** laisse au rédacteur beaucoup plus d'autonomie. Le principe de rédaction est le suivant : lorsque l'utilisateur choisit le type d'information qu'il veut saisir, l'interface affiche les structures syntaxiques qui lui sont associées. L'affichage est fait de manière dynamique, en fonction des choix de l'utilisateur lors du remplissage des cases proposées par le système. Les structures elles-mêmes sont prédéfinies en fonction du verbe choisi par l'utilisateur, tout en sachant que la LC LiSe pose une restriction sur les modifieurs de la phrase : en plus de la négation, une phrase ne peut avoir plus de deux compléments circonstanciels[21] (Figure 3). Chaque structure affiche les éléments obligatoires, bloquants si non-fournis (en gris sur la Figure 3) et les éléments optionnels, non-bloquant (en vert).

Figure 3 Structure de la phrase en LC LiSe

Ce principe de structuration de la phrase autour d'un verbe provient directement de la théorie de dépendances syntaxiques, développé d'abord par L. Tesnière, puis reprise par la suite par de très nombreux chercheurs.

L'interface elle-même est divisée en quatre espaces de travail (Figure 4).

[20] Cette section est basée sur (RENAHY et THOMAS 2009).

[21] Les dénominations telles que complément circonstanciel, sujet, objet etc. ont été choisis à dessin pour correspondre au savoir grammatical tel qu'il a été appris à l'école primaire et secondaire.

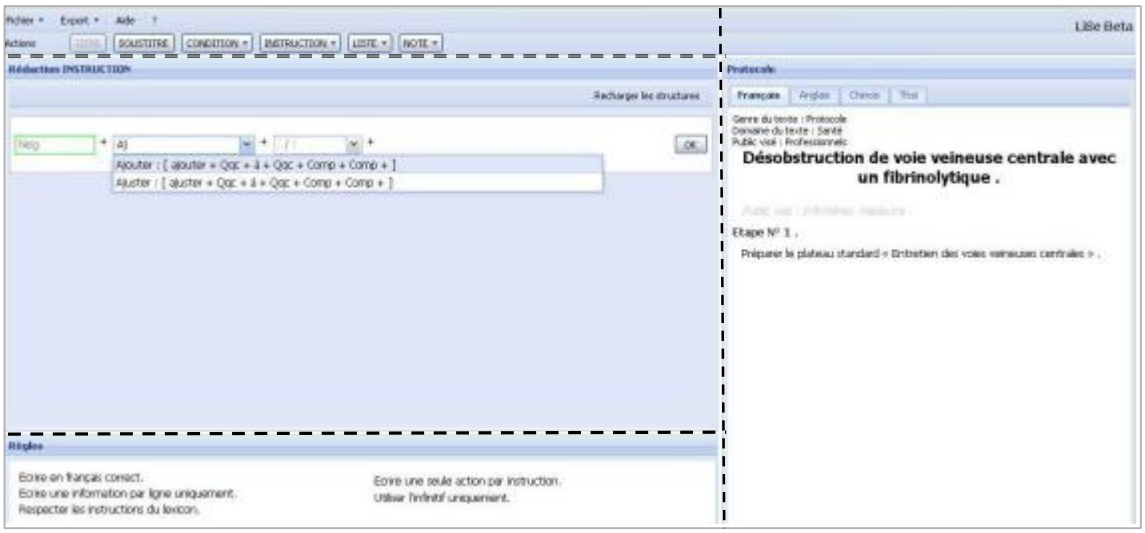


Figure 4 Interface de 'Compagnon LiSe' avec 4 espaces de travail

(1) En haut à gauche : Barre de menu (2) Intermédiaire à gauche : Espace de rédaction (3) En bas à gauche : Espace d'affichage de règles (4) A droite : Espace d'affichage de résultats (en français, anglais, chinois et thaï)

Concrètement, l'**espace de rédaction** permet à l'utilisateur de choisir, à partir de la barre de menu entre les sept structures fonctionnelles prédéfinies pour la LC LiSe et d'afficher les structures syntaxiques qui leurs sont associées. Pour n'en donner qu'un exemple, la rédaction d'un protocole commence obligatoirement avec l'écriture d'un titre (Figure 5). Le rédacteur doit choisir entre 4 structures syntaxiques proposées et jusqu'à ce qu'il n'effectue pas cette action, aucune autre option n'est disponible.

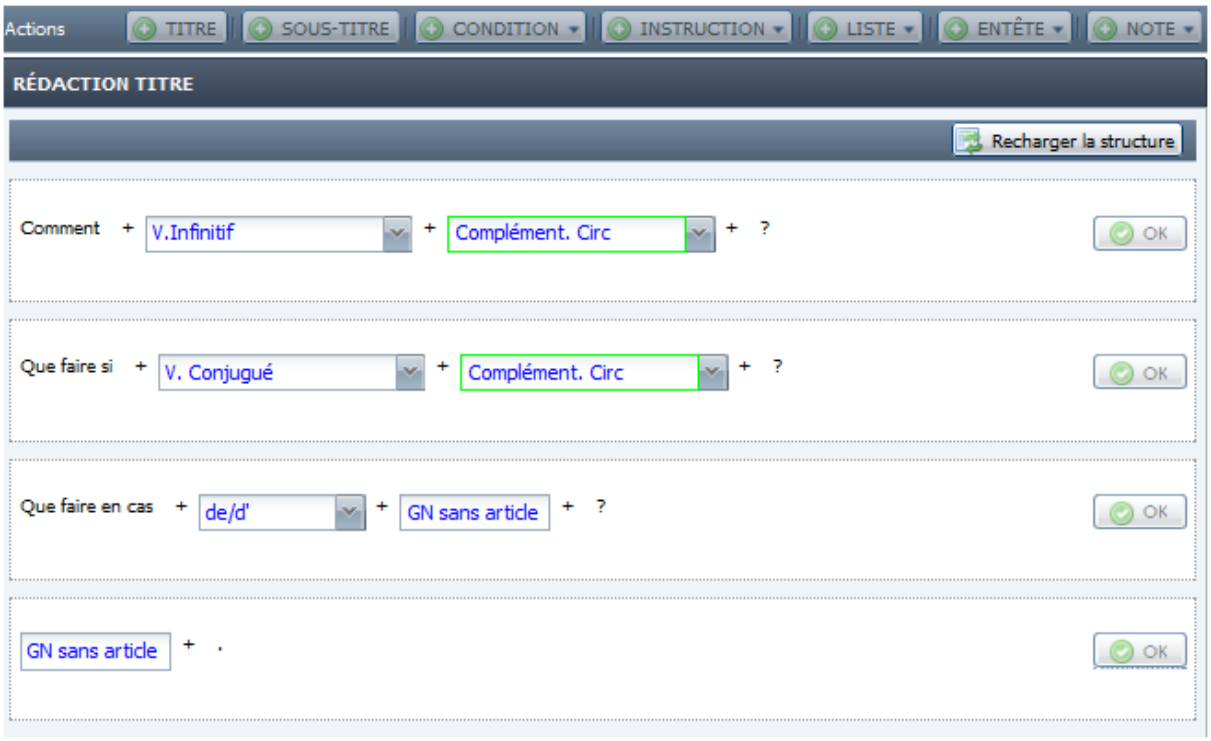


Figure 5 Compagnon LiSe : Affichage des structures associées à un titre.

Les structures syntaxiques prédéfinies peuvent être composées d'éléments très divers, appartenant à plusieurs niveaux d'analyse linguistique, à savoir :

- de signes de ponctuation (« ? », « : ») ;
- d'une ou plusieurs unités lexicales fixes (c'est-à-dire sans qu'on puisse les modifier d'une quelconque façon) et obligatoires (« Que faire », « Comment », « Que faire en cas ») ;
- d'une ou plusieurs unités lexicales obligatoires, mais dont on doit choisir la forme (« de/ d' ») ;
- de catégories syntaxiques, qui doivent être remplacées par des éléments lexicaux (« V. Conjugué », « V. Infinitif ») ;
- de groupes syntaxiques (« GN sans article »), eux aussi remplaçables par des éléments lexicaux ;
- de groupes fonctionnels (« Complément Circ. ») ;
- mais aussi de classes sémantiques générales (« Qqn. », « Qqc. ») ou particuliers (« symptôme », « maladie »), qui constituent les arguments de verbes.

Chaque élément est délimité par une case, qui peut être plus ou moins 'étendu' en fonction des éléments choisis par le rédacteur. Par exemple, le choix d'un verbe particulier entraine l'affichage de son environnement syntaxique (Figure 6). Si l'utilisateur choisit le verbe *distribuer*, l'interface affiche trois cases supplémentaires pour les deux compléments (Qqc, Qqn) et la préposition *à*.

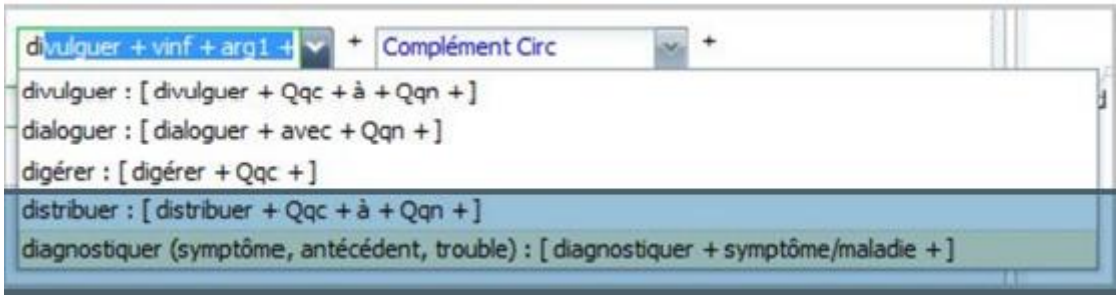


Figure 6 Compagnon LiSe : Exemples de structures verbales.

Dans les faits, la forme finale d'une structure syntaxique dépend des choix que fait l'utilisateur en remplissant les cases. Par conséquent, les quatre structures de base pour rédiger un titre peuvent prendre théoriquement une infinité de formes finales[22].

L'interface aide aussi l'utilisateur à faire les bons choix lexicaux et à utiliser les mots appropriés pour respecter la règle « une unité lexicale = un sens ». D'une part la structure verbale proposée a pour objectif de désambigüiser le sens d'un verbe et de définir son contexte d'utilisation. Si l'on considère le verbe *appliquer*, il ne peut être utilisé que dans la forme *appliquer qqch sur qqch* (cf. *appliquer une pommade sur la plaie*) (Figure 7) et non, par exemple *appliquer qqch à qqch* (cf. *appliquer une méthode d'analyse à une exploration du corpus*).

[22] Limitées tout de même par le nombre de structures verbales et autres, prédéfinies dans la LC.

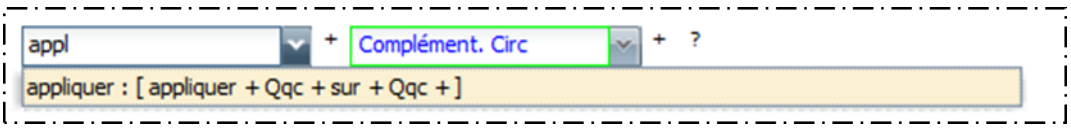


Figure 7 Compagnon LiSe : Structure désambigüisant d'un verbe

Si la structure syntaxique s'avère insuffisante, le sens du verbe est rappelé à l'utilisateur (Figure 8).

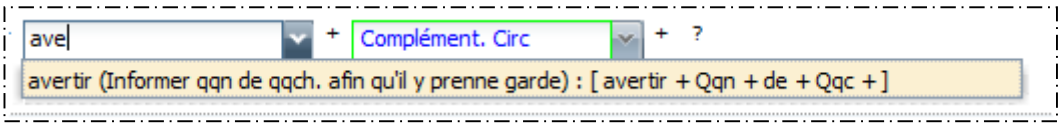


Figure 8 Compagnon LiSe : Définition associée à une structure verbale

D'autre part, un réseau de synonymes associé à un concept d'unités lexicales autorisés et interdites est développé pour prévenir la démultiplication des mots pour le même sens.

Exemple
Dans le domaine de l''immunobiologie, les trois mots suivants sont employés indifféremment (et à tort) pour désigner le même concept :
Calibrage, calibration, étalonnage

Une fois toutes les cases remplies, la phrase complète s'affiche dans **l'espace d'affichage de résultats**, mise en forme selon les formats prédéfinis pour chaque structure fonctionnelle. Elle est également automatiquement traduite et les traductions peuvent être visualisées en cliquant sur les boutons associés aux langues.

Un certain nombre des règles établies pour la LC ont été **entièrement automatisées**. Il s'agit principalement des règles de ponctuation et de mise en forme, qui associent à chaque structure une mise en page spécifique, permettant ainsi une meilleure lisibilité des protocoles. D'autres règles sont **semi-automatiques**, dans la mesure où elles requièrent le choix de l'utilisateur parmi le nombre fini de propositions affichées dans l'interface. Cependant, il est impossible d'automatiser toutes les règles, dont certaines qui sont indispensables pour la rédaction (par exemple : *Écrire 1 idée par phrase ; Respecter l'ordre chronologique* etc.). Nous avons donc dédié un **espace à l'affichage de ces règles** pour les rappeler au rédacteur lors du processus de l'écriture. Pour des raisons d'ergonomie, l'affichage est fait de façon contextuelle, c'est-à-dire que l'on propose seulement les règles pertinentes à la portion du protocole qui est en cours de rédaction. De cette façon, on s'assure aussi que l'écran n'est pas surchargé d'informations non nécessaires à l'utilisateur.

3.3.2 Avantages, inconvénients et perspectives pour le Compagnon LiSe

Le Compagnon LiSe n'est pas un éditeur de texte classique, puisqu'il ne permet pas une saisie libre. Il laisse très peu d'autonomie à l'utilisateur pour s'assurer que celui-ci rédige selon les principes d'une LC. Rappelons qu'il a été conçu dans une double perspective : aider le rédacteur lors du processus d'écriture, mais aussi fournir au système les éléments nécessaires pour la traduction automatique des protocoles. Dans cette seconde perspective,

le rédacteur est impliqué, même s'il n'en a pas vraiment conscience, dans la définition et la délimitation des éléments nécessaire à la traduction. Le système de traduction automatique s'appuie sur les choix faits par le rédacteur lors du guidage : ils permettent de récupérer, désambiguïser et structurer des informations syntaxiques, lexicales et morphologiques en arrière-plan.

Du fait de ces deux fonctions le Compagnon LiSe permet de résoudre la plupart des difficultés liées à la fois à la rédaction en LC et à la traduction automatique, entre autres :

- difficultés liées à la structuration d'information par le système des structures fonctionnelles ;
- difficultés liées à la délimitation et à la désambigüisation automatique des éléments syntaxiques (par exemple, la délimitation entre un argument et un circonstanciel, entre un complément du nom et un complément de la phrase) par le système de cases associés à des éléments syntaxiques ;

Exemple
Evacuer | l'eau de pluie.
V Arg1
Evacuer | l'eau | de | la pièce.
V Arg1 Prep Arg2

- difficultés liées aux choix lexicaux, à la polysémie et à l'homonymie, par la description contextuelle de structures associés aux unités lexicales, mais aussi par un système de réseaux de synonymes et le concept de lexies autorisés et interdites (Figure 9) ;

Exemple

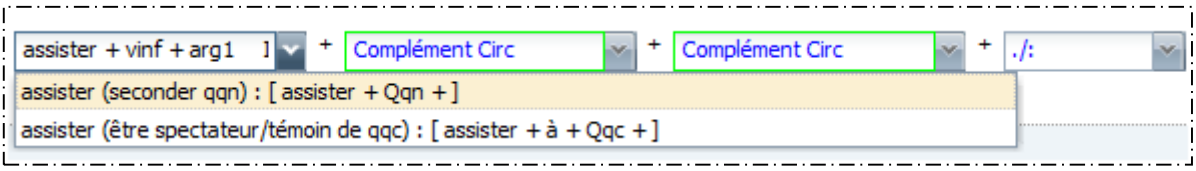


Figure 9 Compagnon LiSe : Structures désambiguïsant du verbe 'assister'

- difficultés liées à l'apprentissage et à la mémorisation des règles de contrôle, par le système de l'écriture prédictive, mais aussi par un système d'étiquetage semi-intuitive (Figure10) ;

Exemple

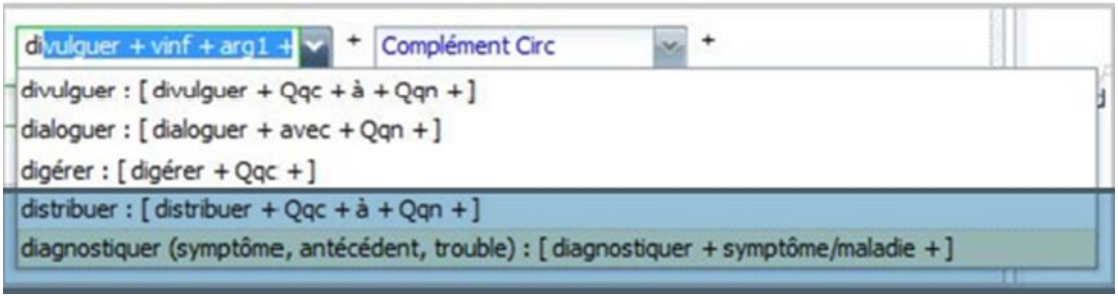


Figure 10 Compagnon LiSe : Ecriture prédictive et étiquetage semi-intuitive

- autres difficultés liées à la traduction, par exemple le choix de la préposition (Figure11).

Exemple

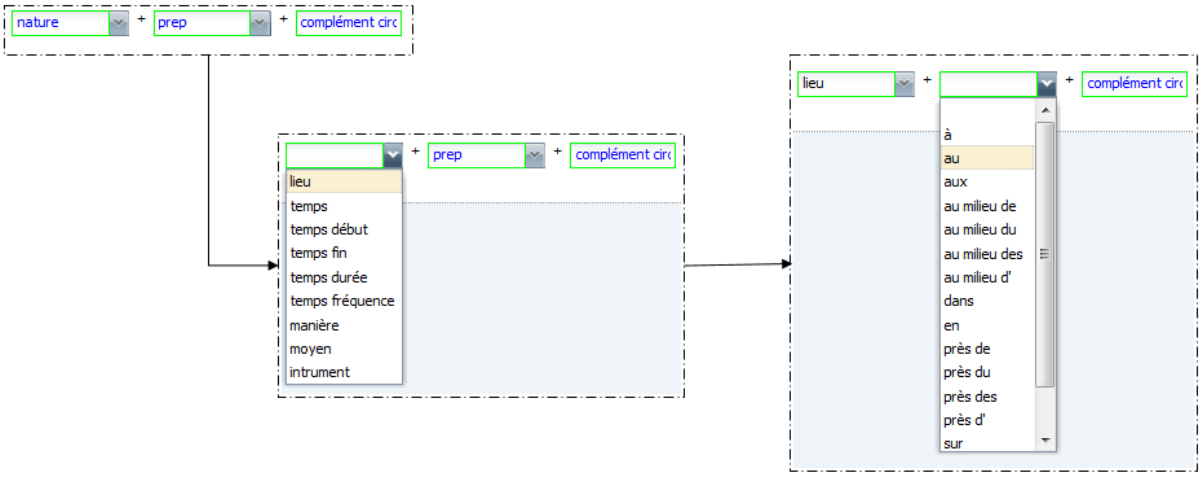


Figure 11 Compagnon LiSe : Choix d'une préposition pour l'introduction d'un complément circonstanciel du lieu

L'ensemble des aides mises en place dans le Compagnon LiSe, permet de répondre, du moins partiellement, à trois critiques majeures à l'encontre de la rédaction en LC évoquées par POWER, SCOTT and HARTLEY (2003), à savoir :

- la nécessité d'entrainer les rédacteurs à l'écriture en LC ;
- la difficulté qu'éprouvent les rédacteurs à trouver la formulation acceptable par le système d'aide à la rédaction ;
- le fait que la sortie d'un système de traduction automatique n'est pas exempt d'erreurs, même si l'entrée est rédigée en LC.

Cependant, ce gain en qualité de rédaction et de traduction a un prix : il est atteint par un contrôle qui nécessite en arrière-plan un travail important d'établissement d'une LC. Une LC avec ce niveau de contrôle doit être établie en fonction du domaine, du moins en ce qui concerne le lexique, mais aussi pour rendre compte de certaines structures syntaxiques spécifiques. Nous avons donc forgé le concept d'une LC 'sur mesure'.

3.4 Langues contrôlées sur mesure : la problématique du lexique

Bien que le principe de LC soit considéré comme bénéfique depuis un certain nombre d'années, le temps nécessaire à sa construction a été estimé par Jeff Allen (2005) de 5 à 10 ans pour des projets industriels, et son coût de production trop élevé pour prétendre à un retour sur investissement rapide (MUEGGE 2009) et même raisonnable. Jusqu'aujourd'hui, cette solution n'était donc envisageable que pour les grandes industries ayant des moyens conséquents (essentiellement l'aéronautique). Or les conclusions de (VUITTON et al. 2009) donnent à penser qu'elles peuvent désormais être également utiles pour les entités plus petites, notamment dans les établissements de santé. La possibilité de concevoir des LC 'sur mesure' en restreignant leur champ d'application à un domaine limité et pour un type de texte

particulier (en d'autres termes, des LC circonscrites) pourrait faire croître de manière exponentielle le rapport coût-bénéfice d'une LC.

Pour répondre à ce besoin, nous avons cherché 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, à un public défini et à 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 (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 ses principes.

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 (2014 : 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 (2014), 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 que nous avons entrepris dans le cadre du projet Sensunique.

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)[23]. Il doit, par ailleurs, prendre en considération le coût et temps de conception, d'où l'idée d'accompagner l'établissement d'une LC 'sur mesure' par des outils automatisés.

Un des obstacles à surmonter concerne le recensement du lexique d'une LC 'sur mesure'. La spécificité du lexique d'une LC 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 comme permises dans le dictionnaire de la LC. De plus, cette contrainte d'exhaustivité du niveau lexical d'une LC implique la distinction d'au

[23] Le cadre de cette collaboration a été défini par Lucie LAROCHE dans le mémoire de Master (2012) que j'ai dirigé : *Méthodologie d'établissement d'une langue contrôlée : application d'une langue contrôlée généralisante à un domaine spécifique.*

moins deux types de dictionnaires : un dictionnaire du lexique d'une LC et un dictionnaire des structures lexicales[24], deux notions que nous allons préciser par la suite.

Ce recensement nous semble possible à condition de disposer d'un corpus délimité et de mettre en place une interaction constante entre un linguiste-terminologue et un spécialiste du domaine afin de réunir les 2 types de compétences complémentaires : linguistiques et métier. Pour rendre le processus d'établissement du lexique plus efficace, nous proposons 3 étapes successives :

- Étape 1 : Une plateforme d'acquisition assistée de vocabulaire (orientée LC) calcule, à partir d'un corpus textuel, un lexique composé d'unités terminologiques (T) et non-terminologiques (NT) priorisées en fonction de leur statut (T ou NT) et de leur potentiel terminologique ;

- Étape 2 : Un premier filtrage sur le lexique est opéré par le linguiste-terminologue pour ne retenir que les unités (terminologiques et non-terminologiques) potentiellement valables ;

- Étape 3 : Un second filtrage aboutissant à la validation finale des unités retenues est réalisé avec l'aide de l'expert du domaine.

La Station Sensunique, une plateforme Web modulaire et collaborative d'aide à l'établissement du lexique d'une LC a été conçue dans l'objectif de faciliter le recensement du lexique d'une LC. Mais avant de décrire ses principes et son fonctionnement, définissons ce que nous entendons par Lexique d'une Langue Contrôlée.

3.4.1 Lexique d'une Langue Contrôlée et Vocabulaire Contrôlé

La notion de lexique d'une LC telle que nous la considérons mérite quelque précisions dans la mesure où elle ne correspond pas tout à fait à la définition du 'vocabulaire contrôlé' ('controlled vocabulary', 'lexique contrôlé', 'termes normalisés') communément employée dans la communauté scientifique, aussi bien francophone qu'anglo-saxonne. Le vocabulaire contrôlé est défini comme « *un langage artificiel utilisé pour classifier et décrire l'information. Il permet la génération de représentations formelles de documents et améliore le repérage de l'information* »[25]. Cette définition du vocabulaire contrôlé reflète l'usage qui en est fait dans le domaine de la recherche d'information (extraction, indexation etc.). Pour ne pas alimenter la confusion entre les deux termes ('vocabulaire contrôlé' et 'lexique contrôlé'), nous proposons de remplacer ce dernier par le terme du Lexique d'une Langue Contrôlée (LLC) et de montrer en quoi ces deux notions diffèrent.

La première différence entre un LLC et un vocabulaire contrôlé vient de leurs objectifs respectifs. Comme le montre la définition précédente, un vocabulaire contrôlé est défini, dans la majorité des cas, pour l'indexation de documents dans le but d'en faciliter la recherche. Par

[24] La notion de structure lexicale n'est pas à confondre avec la notion de structure syntaxique, de structure phrastique, et de structure fonctionnelle etc. (RENAHY et al. 2009).

[25] Source : « NCTTI 39: Normes de l'information et de la technologie du Conseil du Trésor, Partie 2 : Norme du vocabulaire contrôlé » accessible ici : http://www.tbs-sct.gc.ca/pol/doc-fra.aspx?id=15765&section=text

exemple, le MeSH[26], considéré comme vocabulaire contrôlé (NEVEOL 2004), sert à l'indexation de ressources de santé.

La deuxième différence vient de leurs périmètres respectifs. L'ensemble des unités composant un vocabulaire contrôlé renvoie uniquement aux concepts spécifiques d'un domaine. Si l'on admet que ces concepts sont dénommés par des termes, l'objectif est alors de recenser les termes d'un domaine (que ce soit dans des thesaurus, terminologies, onto-terminologies, etc.). 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 couvrir tous les mots d'un texte 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. Sans contredire ces auteurs, nous choisissons de considérer comme *unités lexicales* (UL) toutes les unités d'un LLC.

Pour recenser l'ensemble du vocabulaire contenu dans une collection de textes techniques, plusieurs types de vocabulaire 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 d'immunobiologie, par exemple, plusieurs types de vocabulaire sont nécessaires, allant du vocabulaire spécialisé du domaine jusqu'au vocabulaire général. Par conséquent, y sont nécessairement inclus différents types de lexies :

- les lexies spécialisées du domaine (simples et complexes) :
  - nominaux : *anticorps monoclonaux, réactif de lyse, tampon de fixation* ;
  - verbaux : *numéroter (les cellules), centrifuger (la suspension cellulaire) ;*
  - adjectivaux : *aneuploïde, mononucléé ;*
- les lexies spécialisées d'un autre domaine (*fenêtres informatiques, répartitions gaussiennes*) ;
- les unités du lexique général :
  - soit entrant dans la composition des lexies spécialisées (*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).

Pour répondre aux exigences de non-ambiguïté et de non-redondance, c'est-à-dire au fait qu'une unité lexicale ne peut avoir qu'une seule définition, et qu'une définition ne peut correspondre qu'à une seule unité lexicale dans un domaine choisi, il s'avère nécessaire de contrôler l'ensemble du lexique utilisé pour la conception de la documentation dans un

[26] Le MeSH (Medical Subject Headings), thésaurus de référence dans le domaine biomédical, http://mesh.inserm.fr/mesh/.

domaine (pour éviter, par exemple d'employer le mot *solution* au sens général dans les protocoles d'immunobiologie, dans lesquels *solution* prend un sens très spécifique[27]), ainsi que recueillir les informations complémentaires concernant les liens qu'entretiennent les différentes unités entre elles (synonymie, inclusion, variation etc.).

Pour respecter la contrainte de non-ambigüité, il s'agit de recenser les cas de polysémie afin de restreindre la signification de chaque UL à une seule acception.

Pour répondre à l'exigence inverse de non-redondance, il s'agit de contrôler l'ensemble des relations paradigmatiques d'une UL et recenser différents types d'informations complémentaires relatives à la variation : synonymie (*calibration, calibrage*[28]), variation orthographique, acronymes et abréviations (*anticorps monoclonal, AcMo*), variation morphologique flexionnelle et dérivationnelle (*type de cellules, type cellulaire*), variation morphosyntaxique (tels que les phénomènes d'élision de mots vides (*isotype de contrôle, isotype contrôle*), la permutation, la coordination (*protocole de lyse et lavage*). Certains exemples vont cumuler plusieurs de ces phénomènes linguistiques, comme dans *histogramme des phénotypes B27 positifs et négatifs, contrôles HLA-B27 positif et négatif, échantillons HLA-B27 négatifs, échantillon B27 négatif*.

La tâche est ardue : pour empêcher la redondance, il ne suffit pas de recenser les unités lexicales permises, mais aussi celles qui ne doivent pas être utilisées. Ainsi, un rédacteur ne se retrouvera pas bloqué lors du processus d'écriture : toutes les UL recensées comme interdites pointeront vers une UL autorisée, permettant alors à un outil d'aide à la rédaction d'alerter le rédacteur de la non-conformité de l'UL et de lui proposer un substitut.

3.4.2 Structures lexicales

Nous introduisons la notion de structure lexicale pour répondre aux besoins d'exhaustivité du recensement du lexique d'une LC. Nous nommons Structures Lexicales (SL) un patron morphosyntaxique imposé et contrôlé par un lexème avec une partie fixe et une partie variable. Ce lexème est prédicatif, sauf dans quelques cas particuliers que nous mentionnerons plus tard. Il s'agit donc pour la plupart de structures verbales, nominales ou adjectivales (et potentiellement adverbiales) gouvernées par un lexème particulier (respectivement verbe, nom ou adjectif et potentiellement adverbe). On crée des structures lexicales lorsqu'une suite, dépendante d'un lexème, présente une certaine variabilité et, par la même, est impossible à encoder dans un dictionnaire de lexies spécialisées.

Exemple
marquage des cellules
marquage des cellules leucocytaires
marquage des cellules endothéliales vasculaires animales
marquage des cellules souches
marquage des cellules en suspension
etc.

[27] « *Liquide formé par la dissolution d'une substance solide (p. ex. médicament) dans un solvant* », GDT.
[28] Bien que *calibrage* soit un terme français et *calibration* anglais, les rédacteurs les utilisent comme de vrais synonymes.

Puisqu'il est difficile de prévoir toutes les suites introduites par le lexème 'marquage', nous proposons de l'encoder dans un dictionnaire spécifique, sous forme d'une structure en partie lexicalisée :

« marquage de < NOM : CELLULE> »

Dans cette suite, les chevrons (<>) introduisent la partie variable, souvent définie par sa catégorie fonctionnelle (ici: NOM, une suite de catégories grammaticales en fonction du groupe nominal), qui peut être en plus caractérisée par son appartenance à une classe sémantique (ici : CELLULE).

La plupart des structures lexicales sont contrôlées par des lexèmes prédicatifs, tels que les verbes (« marquer < NOM : CELLULE> »), les noms déverbaux (« marquage de < NOM : CELLULE> »), les participes en fonction adjectivale (« < NOM : CELLULE> marqué(es) > »). Il arrive cependant qu'on crée une structure lexicale pour exprimer un patron non-prédicatif, mais qui est utilisé fréquemment dans un domaine et un corpus donné :

Exemple
(anticorps <(ADJ:OBTENTION)* ADJ:TYPE-ANTICORPS)* (humain)|(de NOM:ANIMAL))*
anticorps Fab'2 de chèvre
anticorps primaire de souris

La notion de structure lexicale est primordiale lorsque, en s'éloignant de la théorie terminologique classique, nous voulons considérer comme termes d'autres syntagmes que les syntagmes nominaux. En effet, certains verbes ou adjectifs peuvent renvoyer à des notions bien spécifiques dans des domaines précis. Certains dictionnaires terminologiques recensent d'ores et déjà des termes de nature verbale ; c'est le cas, par exemple, du Grand Dictionnaire Terminologique (GDT)[29], dans lequel on trouve aussi bien le nom 'centrifugation' que le verbe 'centrifuger'. Par contre, la description de ce verbe, s'arrêtant à l'identification de sa catégorie verbale, paraît incomplète : 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 complété par des compléments d'une certaine classe sémantique :

« centrifuger <NOM : SANG> »

Un autre avantage concernant l'encodage en structures lexicales consiste à établir des relations entre les différents sens des lexies dérivées et à vérifier la cohérence du recensement du vocabulaire. Normalement, les lexies en relation de dérivation ne peuvent introduire dans leurs structures que des arguments appartenant à des classes sémantiques identiques :

Exemple
« numéroter < NOM : CELLULE> » ; « <NOM : CELLULE> numéroté(es) » ; « numération de <NOM : CELLULE> »
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

[29] Banque de données terminologiques élaborée par l'Office québécois de la langue française : http://www.granddictionnaire.com/

sémantique *CELLULE* et que *numéroter, numération* (voire le participe passé adjectival *numéroté*) renvoient toujours à la même classe sémantique.

Les SL, telles que nous les entendons, ont donc, pour tout élément prédicatif (verbe, nom, adjectif et parfois même adverbe) des propriétés distributionnelles et des propriétés conceptuelles de sous-catégorisation.

Les SL sont à mi-chemin entre le lexique et les règles syntaxiques d'une LC. Le point de rencontre étant la présence d'un lexème régissant une structure. Elles se rapprochent des grammaires lexicalisées dans la mesure où ces dernières donnent un rôle important au lexique dans la description de la langue : Le Lexique-Grammaire de M. GROSS, la Lexicologie Explicative et Combinatoire de la Théorie Sens-Texte de MEL'CUK, les projets de type FrameNet[30] basés sur la notion de sémantique des cadres, ou VerbNet[31] recensent aussi le lexique avec les mêmes objectifs. Dans ce type de grammaire, le lexique est recensé avec l'ensemble de ses propriétés syntaxico-sémantiques.

3.4.3 Informatisation du processus du recensement du Lexique d'une Langue Contrôlée : la Station Sensunique

La Station Sensunique a été conçue dans l'objectif initial d'assister le processus de constitution du lexique d'une LC telle que définie dans la section précédente, et de diminuer le temps (et donc le coût) nécessaire à sa conception. Automatiser ce processus impliquait deux types de contraintes : (1) liées au recensement du lexique spécialisé à partir d'un corpus et (2) liées à la conception d'une LC, 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 présupposait la gestion des 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) (Figure 12).

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 possibles d'utilisation des ressources à construire, en termes de : domaines, types de textes, public cible 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* (CHIARCOS et al. 2012)). Elle offre aussi les fonctionnalités adéquates aux étapes suivantes du processus: les deux premières sont la sélection et la validation des UL par un analyste, et la seconde est la validation par l'expert métier et export de la ressource finale exploitable. Les interfaces utilisateur (interface de gestion et interface de travail), ne nécessitent aucun savoir-faire technique et sont faciles à prendre en main et à explorer. L'application exploitant les

[30] https://framenet.icsi.berkeley.edu/fndrupal
[31] http://verbs.colorado.edu/verb-index/

lexiques d'une LC à concevoir (besoin initial de la Station) est le 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).

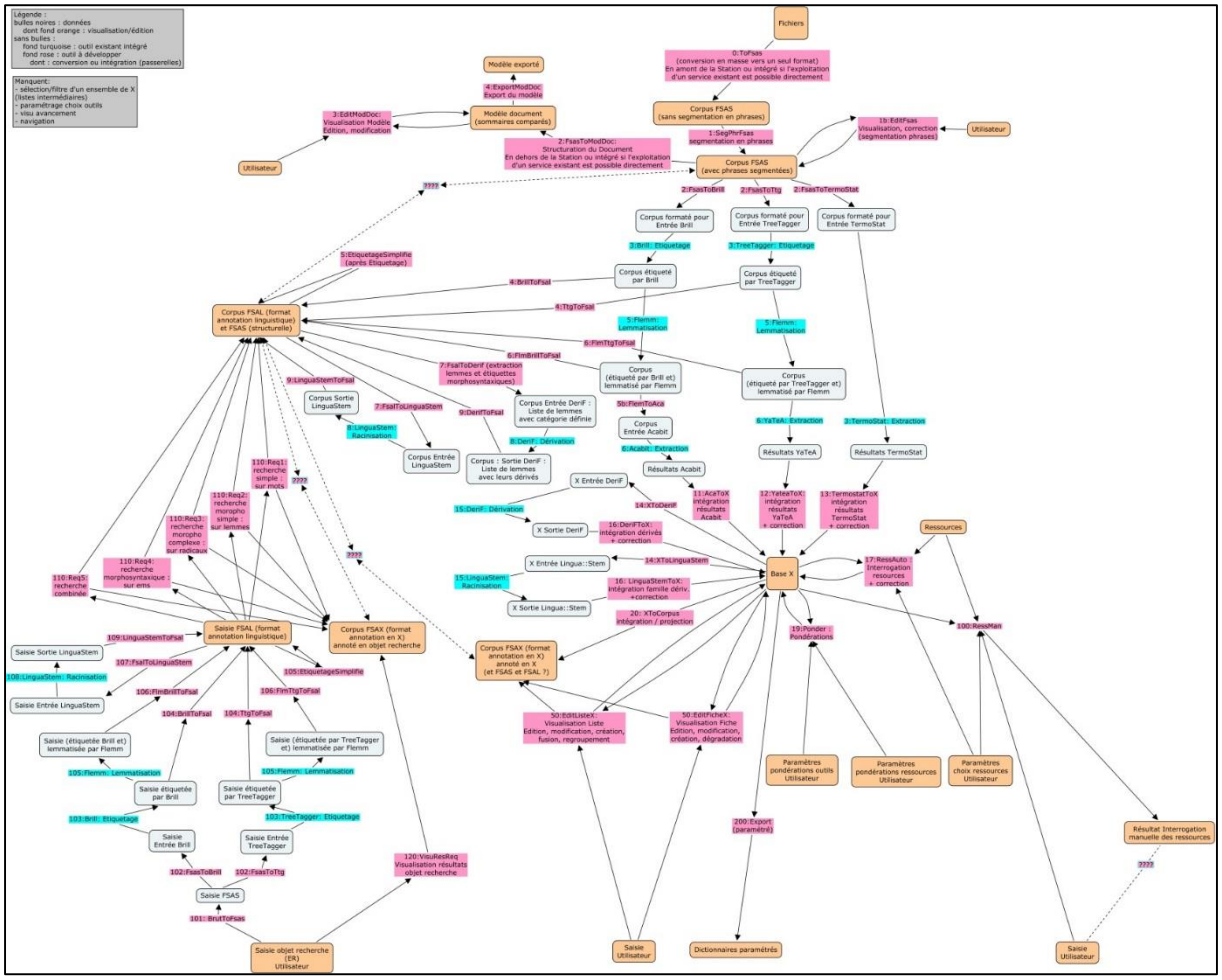


Figure 12 Schéma de la Station Sensunique du point de vue de ses fonctionnalités.

3.4.3.1 État de l'art des outils d'extraction et de gestion terminologique

Puisqu'aucun outil spécifique au recensement d'un LLC n'existe à ce jour, nous avons vérifié dans quelle mesure les extracteurs terminologiques peuvent aider à accomplir cette tâche. Leur objectif est clairement différent (ils visent uniquement les termes). Cependant, de par leur fonction d'extracteur de vocabulaire terminologique, ils se rapprochent le plus de nos besoins.

Les outils d'extraction ou d'acquisition terminologique ont été développés pour des applications diverses telles que la recherche d'information, l'extraction d'information, la veille ou l'acquisition d'ontologies. Ils reposent sur deux types de méthode : statistique et linguistique. Les synthèses très complètes de ces méthodes ont été proposées par divers auteurs (BOURIGAULT et al. 2000 ; DROUIN 2003 ; L'HOMME 2004). Il ressort de la plupart de ces travaux que les outils hybrides exploitant les deux méthodes sont les plus performants pour les corpus spécialisés. Nous ne listerons pas non plus la totalité de ces outils ici, puisqu'il

existe déjà de multiples publications le faisant dont, notamment (CABRE et al. 2001) mais en citons quelques-uns utilisables sur le français, par ordre alphabétique : ANA (ENGUEHARD 1992), Acabit (DAILLE 1994), EXIT (ROCHE et al. 2004), Lexter (BOURIGAULT 1994), TermoStat (DROUIN 2002), YaTeA (AUBIN et al. 2006).

La méthode la plus simple et la plus utilisée repose sur la répétition de formes : les unités lexicales (ou leurs matrices morphosyntaxiques) les plus répétées sont les plus représentatives (on joue alors sur la fréquence). L'association entre deux mots sert à savoir s'ils composent un terme potentiel ou un candidat terme (CT). Enfin, l'opposition de corpus permet de faire ressortir le vocabulaire spécifique d'un corpus par rapport à un autre.

Divers indices ou mesures statistiques découlent de ces approches et certains ont été implémentés dans les logiciels d'acquisition automatique de termes dans le but premier de réduire le bruit présent dans leurs résultats, de pouvoir filtrer ou classer les CTs. Nous pouvons citer tout d'abord la fréquence (brute ou relative) dont l'intérêt pour l'évaluation du potentiel terminologique d'un CT dans les corpus spécialisés a été démontrée (DAILLE et al. 1994). Le calcul de spécificité (LAFON 1980) conçu pour cerner le vocabulaire spécifique à un sous-corpus par rapport à l'ensemble d'un corpus et adapté (DROUIN 2002) grâce à une fusion d'un corpus de référence et d'un corpus d'analyse afin de vérifier si le lexique de ce dernier se comporte comme le lexique du premier. Le test du X^2 utilisé pour l'analyse des conversations au sein du British National Corpus (RAYSON et al. 1997), pour évaluer l'homogénéité des corpus (KILGARRIFF 2001) ou pour comparer les fréquences d'occurrence de CT (DROUIN 2002). Le rapport de vraisemblance, ou *log-likelihood*, (DUNNING 1993) utilisé, entre autres, pour la comparaison de corpus (RAYSON et al. 2000) sert également à faire ressortir les meilleurs CT (DROUIN 2002). Le log-odds ratio exploité pour les collocations (EVERT 2004) permet de dire qu'un CT est potentiellement intéressant d'un point de vue terminologique. L'information mutuelle (CHURCH et HANKS 1990) calcule une certaine forme d'indépendance de 2 mots d'un terme complexe et tend ainsi à faire ressortir les termes rares et peu fréquents.

Il nous faut également citer, sur un autre plan (non statistique) l'exploitation de ressources terminologiques existantes, une méthode dite exogène (HAMON 2006) permettant de filtrer les CT et implémenté dans le logiciel YaTeA.

Les extracteurs terminologiques sont considérés aujourd'hui matures (CERBAH et al. 2006) mais cette affirmation dépend de l'objectif du recensement du vocabulaire. La maturité des extracteurs est certainement suffisante pour construire les vocabulaires contrôlés, dans le cadre de recherche d'information. Cependant, lorsqu'il s'agit de la construction de dictionnaires spécialisés ou de la traduction, trois problèmes majeurs peuvent être invoqués :

- attribution du statut terminologique à un non-terme ;
- présence d'un bruit trop important dans les résultats ;
- manque de propositions d'unités pertinentes (le silence).

Puisque l'acquisition d'un LLC peut être comparée jusqu'à un certain point à l'acquisition terminologique, nous avons choisi de nous appuyer sur les Extracteurs de Termes (EdT), tout en tentant d'améliorer leurs résultats pour qu'ils répondent à nos besoins. Plus précisément, nous avons évalué le bénéfice que nous pourrions tirer de la coopération de

plusieurs EdT. De multiples travaux fondés sur la coopération d'outils ont démontré son intérêt : en premier pour la reconnaissance vocale avec le système ROVER (FISCUS 1997), repris, pour n'en citer que quelques-uns, pour des analyseurs syntaxiques (BRUNET-MANQUAT 2004) ou des étiqueteurs morphosyntaxiques (SERP 2008). Mais ce principe n'a jamais été appliqué aux EdT.

3.4.3.2 Expériences préliminaires : multi-extraction comme principe de recensement du LLC

Nous avons posé et vérifié l'hypothèse (H1) que l'utilisation simultanée de plusieurs EdT (que nous appellerons désormais la « multi-extraction ») est plus profitable (qu'un seul) au recensement du LLC. Nous avons subdivisé cette hypothèse en deux parties :

> *(H1.1) : Les résultats proposés par plusieurs EdT sont les candidats-termes (CT) les plus pertinents, et peuvent être considérés comme UL terminologiques (la multi-extraction permet de déterminer le statut terminologique d'une UL en faisant ressortir son potentiel terminologique).*
>
> *(H1.2) : Les CT non valides sont des UL candidates non terminologiques potentiellement pertinentes (le bruit des EdT, dans leur fonction initiale de recensement des termes, peut diminuer le silence, dans la fonction détournée de recensement des UL d'un LLC). Si cela est confirmé, la tâche de recensement d'un LLC peut être organisée, en classant les résultats par poids terminologique[32] (Pt) et/ou comme aide au filtrage du bruit pour l'acquisition de lexiques terminologiques.*

L'expérimentation a été effectuée sur un corpus de référence de 14 modes opératoires d'immunobiologie (10 064 mots) de l'Établissement Français du Sang Bourgogne Franche-Comté (EFS B/FC)[33]. Notons que la méthodologie est indépendante du domaine. Nous avons construit manuellement sur la base de ce corpus un LLC de référence, grâce à des critères linguistiques, la consultation de ressources terminologiques[34] et d'experts métier[35]. Le lexique de référence obtenu contient 1 512 UL (lemmes) pour 1 729 formes fléchies (utilisées en corpus), 7 catégories syntaxiques fonctionnelles distinctes (distinction minimale nécessaire pour un LLC : Adjectif, Adverbe, Nom, Nom propre, Verbe au participe passé, Verbe au participe présent, Verbe hors participes), 92 matrices morphosyntaxiques distinctes (exemple : *Nom Prep Det Nom Prep Det Nom Prep Nom* pour *fraction de l'immunoglobuline de l'antisérum de lapin*) et 2 statuts lexico-terminologiques distincts (terminologique et général).

Afin d'estimer l'utilisabilité et l'adéquation technique des EdT à nos besoins et de nous limiter à 3 EdT (coût raisonnable de la tâche d'évaluation), nous avons pris en compte les critères suivants :

- langue : français ;
- méthode : non purement statistique[36] : linguistique ou hybride ;
- disponibilité : de suite ;

[32] Pt est un indice de fiabilité d'une ULC en tant que terme (relatif à son potentiel terminologique).
[33] Documents décrivant le déroulement détaillé et structuré des différentes étapes d'une manipulation.
[34] Le Grand Dictionnaire Terminologique, Termium Plus, le dictionnaire médical Masson 5ème édition.
[35] EFS B/FC, partenaire Santé dans le projet Sensunique.
[36] A cause de la taille estimée des corpus utilisés pour concevoir une LC.

- licence : libre ou commerciale ; dans ce cas, coût faible ou nul ;
- maturité de l'outil : non prototype ;
- environnement informatique : Unix ;
- modalité d'exécution : service web ou appel en ligne de commande ;
- temps d'exécution : respectant le seuil d'appel en web service ;
- et domaine d'application : non spécifique.

Ces critères nous ont menés aux EdT Acabit (DAILLE 1994), TermoStat (DROUIN 2003) et YaTeA (AUBIN et al. 2006). Acabit procède par identification de groupes nominaux complexes sur des matrices syntagmatiques pour extraction de bi-termes, regroupement de variantes (à partir de ces bi-termes) puis filtrage statistique. YaTeA enchaîne identification de groupes nominaux à partir de frontières morphosyntaxiques, calcul de leurs structures en tête et modifieur, puis exploitation de ces structures pour l'analyse des groupes nominaux restants. Enfin, TermoStat fonctionne par détection de CT sur patrons morphosyntaxiques puis pondération et filtrage selon la spécificité de chaque CT (méthode de mise en opposition de corpus spécialisés et non spécialisés). De surcroit, YaTeA et TermoStat ont l'avantage d'extraire des termes simples en plus des termes complexes ; et TermoStat est le seul à extraire également des termes non nominaux.

Deux des tâches du linguiste lors de la conception d'un LLC ont été évaluées :

1. Recensement des UL (de l'ensemble des UL d'un LLC) ;
2. Recensement des termes (des UL de statut terminologique d'un LLC).

Pour chacune de ces tâches, nous avons procédé à 3 expérimentations :

1. Évaluation des résultats de chaque EdT pris séparément ;
2. Évaluation des résultats cumulés de tous les EdT (union) ;
3. Évaluation des résultats consolidés, ou communs (intersection).

Pour chaque évaluation, nous avons calculé les mesures suivantes :

- Précision : P = (formes extraites correctes) / (formes extraites) ;
- Rappel : R = (formes extraites correctes) / (formes de référence) ;

Notre objectif étant d'estimer la capacité des EdT à recenser les UL (terminologiques ou non) et non leur capacité de lemmatisation ou de variation terminologique, nous avons opté pour l'appariement des résultats sur la base de comparaison des formes fléchies extraites et celles du lexique de référence.

Les résultats de la tâche du recensement des UL sont présentés dans le Tableau 1.

Expérimentations	Outil(s)	P	R
Résultats d'un EdT	TermoStat	64 %	40 %
	YaTeA	43 %	**52 %**
	Acabit	44 %	17 %
Résultats cumulés (union)	TermoStat ∪ YaTeA	44 %	68 %
	TermoStat∪ ACABIT	55 %	48 %
	YaTeA ∪ ACABIT	41 %	59 %
	TermoStat ∪ YaTeA ∪ ACABIT	**42 %**	**72 %**
Résultats communs (intersection)	TermoStat ⋂ YaTeA	**74 %**	22 %
	TermoStat ⋂ ACABIT	63 %	9 %
	YaTeA ⋂ ACABIT	62 %	11 %
	(TermoStat ⋂ YaTeA) ou (TermoStat ⋂ ACABIT) ou (YaTeA ⋂ ACABIT)[37]	**69 %**	29 %

Tableau 1 Tâche de recensement des UL

Dans le meilleur des cas, en utilisant un seul EdT, 52 % (valeur en gras, Table 1) des UL du LLC sont recensées, ce qui est loin de satisfaire le critère d'exhaustivité.

Le cumul des résultats des 3 EdT permet de couvrir quasiment ¾ du lexique de référence (rappel de 72 %, en gras, Table 1). Ceci confirme l'hypothèse H1 : la multi-extraction permet de mieux couvrir le LLC que l'utilisation d'un seul EdT. En revanche, dans ce cas, il reste à filtrer manuellement près de 60 % des propositions et il devient nécessaire de filtrer automatiquement le bruit.

La combinaison d'EdT obtenant la meilleure précision est TermoStat + YaTeA (74 %, Table 1). Cependant, il apparaît également que n'importe quelle combinaison de 2 EdT donne une précision de 69 % (donc légèrement plus faible). Nous proposons de filtrer le bruit sur cette dernière combinaison en considérant que ce cas de figure sera plus généralisable (à d'autres domaines) dans la mesure où il « suffit » qu'une UL soit proposée par 2 EdT pour être estimée pertinente. L'opération consisterait à augmenter la valeur d'un indice relatif au potentiel terminologique des UL concernées (proposées par 2 EdT), et creuser ainsi l'écart avec celles qui ne sont pas proposées que par un EdT. Cela revient à distinguer les ULC à fort potentiel terminologique de celles à faible potentiel terminologique en les classant et non en supprimant ces dernières.

Les résultats de la tâche de recensement de termes sont montrés dans le Tableau 2.

[37] Sur l'ensemble des résultats communs à (proposés par) au moins 2 EdT, quels qu'ils soient.

Expérimentations	Outil	P	R
Résultats d'un EdT	TermoStat	28 %	52 %
	YaTeA	16 %	58 %
	ACABIT	14 %	17 %
Résultats cumulés (union)	TermoStat ∪ YaTeA	16 %	76 %
	TermoStat∪ ACABIT	23 %	60 %
	YaTeA ∪ ACABIT	14 %	63 %
	TermoStat ∪ YaTeA ∪ ACABIT	15 %	**79 %**
Résultats communs (intersection)	TermoStat ⋂ YaTeA	**37 %**	33 %
	TermoStat ⋂ ACABIT	24 %	11 %
	YaTeA ⋂ ACABIT	26 %	14 %
	(TermoStat ⋂ YaTeA) ou (TermoStat ⋂ ACABIT) ou (YaTeA ⋂ ACABIT)	31 %	39 %
	TermoStat ⋂ YaTeA ⋂ ACABIT	32 %	9 %

Tableau 2 Tâche de recensement des termes

La mesure de précision de 37 % (TermoStat ⋂ YaTeA, Table 2) pour les résultats communs permet de valider l'hypothèse H1.1. : la multi-extraction aide à déterminer le statut terminologique d'une UL en faisant ressortir son potentiel terminologique. La différence (même faible) de rappel entre les résultats cumulés des 3 EdT pour le recensement des termes (79 %, Table 2) et le recensement des UL (72 %, Table 1) démontre qu'une partie des candidats proposés ne sont pas des termes mais sont, pour le LLC, des UL correctes, de statut non terminologique (hypothèse H1.2). Bien que les résultats soient moindres que ceux escomptés, ils demeurent satisfaisants et il est possible qu'ils soient meilleurs sur des corpus plus conséquents.

En résumé, cumuler les résultats de tous les EdT permet de couvrir 79 % des termes (rappel TermoStat ∪ YaTeA ∪ ACABIT, Table 2), et le meilleur moyen d'aider à déterminer le statut d'une UL est, non pas de se baser sur les résultats communs aux 3 EdT (contrairement à ce que nous attendions), mais de se baser sur les résultats communs aux 2 EdT TermoStat et YaTeA (précision de 37 % dans la Table 2). Ceci valide tout de même l'hypothèse selon laquelle la multi-extraction aide à recenser et à organiser la validation d'un LLC.

Le fait que la multi-extraction permette à la fois de réduire le silence et le bruit des propositions nous incite à introduire un indice, relatif au potentiel terminologique, variable en fonction des résultats des EdT. Nous proposons d'attribuer à chaque UL candidate un poids terminologique Pt,puis de faire varier ce Pt initialement nul en fonction des résultats de chaque EdT. Nous proposons une stratégie de variation du Pt consistant à augmenter du Pt d'une UL en fonction du nombre d'EdT qui la proposent comme candidate. Ce principe traduit bien les faits suivants :

- un EdT propose un candidat « terme », donc un candidat ayant un potentiel terminologique ;
- un candidat a d'autant plus de probabilité d'être un terme fiable qu'il y a d'EdT le proposant (comme candidat) ;
- les résultats pourront être classés et validés selon la valeur du Pt.

Notons que les 3 EdT utilisés intègrent *a priori* (TermoStat) ou *a posteriori* (Acabit et YaTeA) des indices statistiques afin de cerner les termes les plus pertinents. Bien qu'il puisse être intéressant de coupler les valeurs de ces indices (différents pour chaque candidat) au Pt, nous avons fait le choix de ne pas le faire expressément, afin de ne pas rendre l'algorithme de pondération dépendant des EdT utilisés (et puisque le calcul de Pt repose déjà indirectement sur l'efficacité des EdT utilisés).

L'exploitation des résultats des expérimentations menées nous a permis de proposer une méthode d'acquisition d'un LLC et d'optimisation de l'acquisition terminologique. Elle repose sur la coopération de plusieurs EdT et permet de faire ressortir le potentiel terminologique des candidats, de réduire le silence obtenu avec un seul EdT et de filtrer le bruit en classant les candidats sur un indice de potentiel terminologique.

Outre concevoir un outil dédié au recensement d'un LLC, l'originalité de ces travaux réside dans le fait que nous proposons de faire coopérer plusieurs EdT pour améliorer leurs résultats et mettre en place un système de filtrage, alors que les travaux antérieurs d'évaluation d'EdT visaient leur mise en opposition (ou classement) (GRABAR 2004).

Pour améliorer l'extraction terminologique, nous avons mis au point un système à base de vote, sur la méthode dite du « vote à la majorité » (BRUNET-MANQUAT 2004) où plus un terme est proposé par différents EdT, plus sa fiabilité est renforcée.

Nous avons conçu une plateforme implémentant cette méthode. Elle intègre les étiqueteurs morphosyntaxiques Brill[38] et TreeTagger, le lemmatiseur Flemm (NAMER 2000) pour les analyses préalables et nécessaires à l'extraction, et les EdT Acabit, TermoStat et YaTeA. Elle permet de procéder à l'extraction et à l'organisation de lexique terminologique et non-terminologique à partir d'un corpus français au format XML TEI P5. La plateforme est paramétrée par défaut sur le principe du vote à la majorité mais l'utilisateur peut ajuster le poids attribué à chaque EdT, en fonction de ses besoins, afin de rendre cette plateforme aussi flexible que possible. Nous avons également intégré un module d'interrogation de ressources terminologiques ou lexicales existantes, ce qui permet de renforcer, une nouvelle fois, la fiabilité du potentiel terminologique des candidats.

3.4.3.3 Description de la Station Sensunique

Comme toute plateforme terminologique (par exemple : HyperTerm[39], Terminae[40], Terminus[41]), la Station (Figure 13) intègre la mise en séquencement de plusieurs outils TAL (étiquetage, lemmatisation et extraction de termes)[42]. 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 de filtrer automatiquement le bruit. Plus

[38] Avec le lexique et le fichier de règles fournis par l'ATILF-CNRS, de Nancy.
[39] http://www.tedopres.com/hyperterm-terminology-management [03/04/2014].
[40] http://lipn.univ-paris13.fr/terminae/index.php/Main_Page [03/04/2014].
[41] http://terminus.iula.upf.edu//cgi-bin/terminus2.0/terminus.pl [03/04/2014].
[42] Ce paragraphe est basé sur Thomas et al. (2014a et 2014b).

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.

Figure 13 Interface de connexion de la Station Sensunique

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 soit pas nouveau, 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 :

- chronologique (centré processus) : import des textes d'entrées, analyse automatique[43], 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), 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 Sensunique fonctionne de façon modulaire, chaque module proposant à l'utilisateur plusieurs services (cf. Figure 14). Chaque module correspond au processus d'acquisition de ressources, divisé en plusieurs étapes :

- Étape 0 : Création d'un projet et gestion des utilisateurs ;
- Étape 1 : Analyse automatique, qui extrait, à partir d'un corpus, une liste composée d'unités terminologiques et non-terminologiques classées en fonction de leur statut et de leur potentiel terminologique ;
- Étape 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 ;
- Étape 3 : Définition des paramètres d'export et export des ressources établies

[43] L'analyse automatique comprend : étiquetage, lemmatisation, racinisation, extraction des ULC, interrogation des ressources externes et internes, calcul des pondérations.

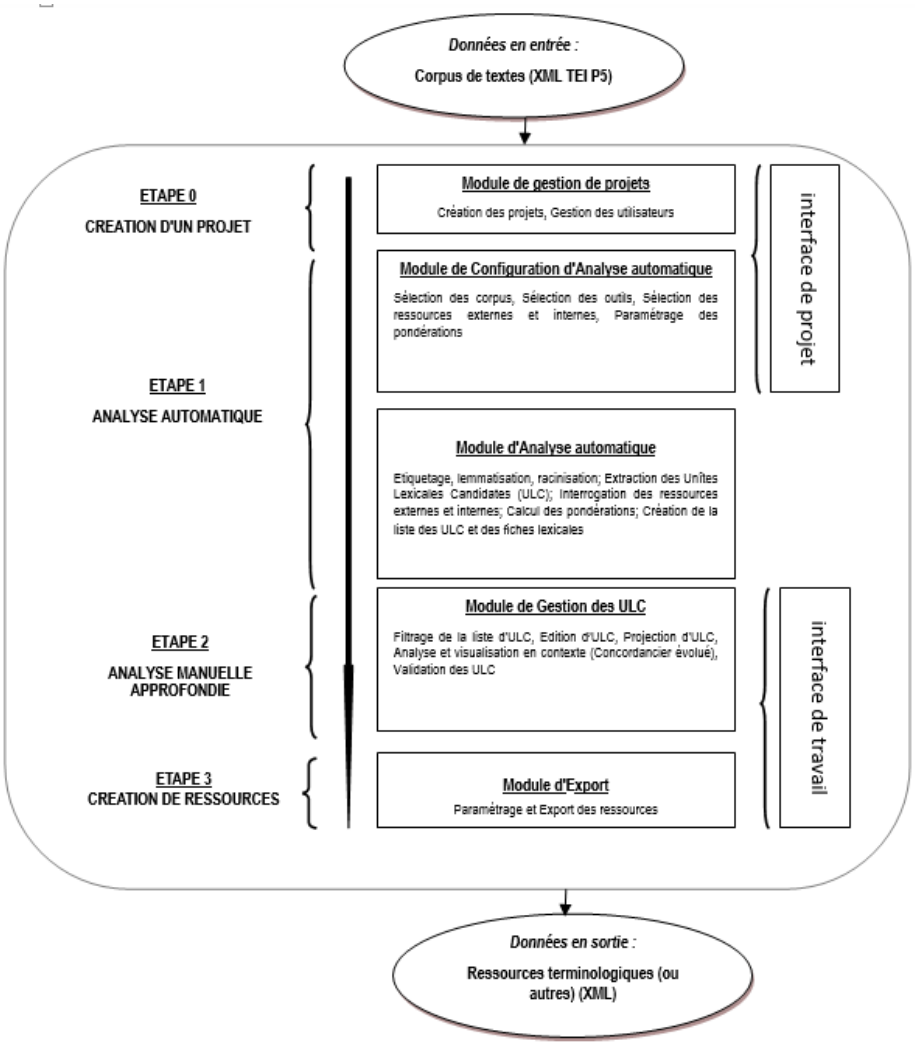


Figure 14 Schéma de la Station Sensunique du point de vue chronologique

3.4.3.3.1 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. La configuration de l'analyse s'effectue à partir de la première interface de la Station Sensunique : interface du projet (Figure 15).

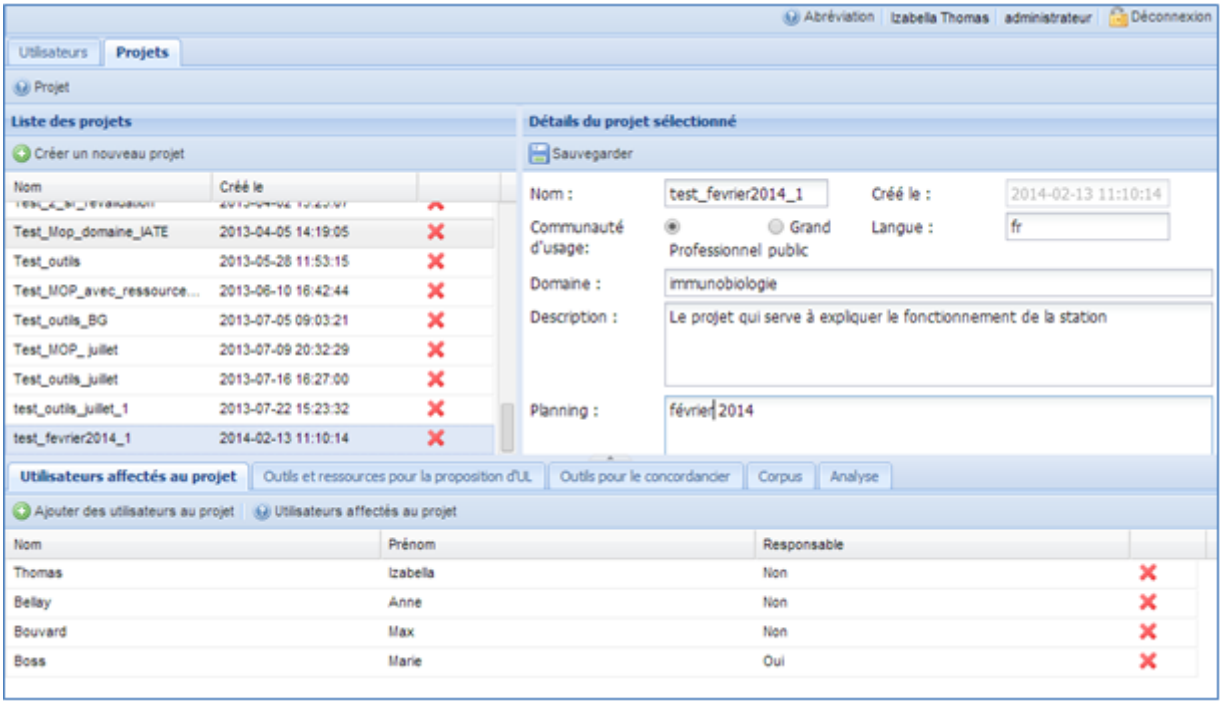


Figure 15 Station Sensunique : Interface du projet (onglet Projet)

Sélection de corpus

Pour le même projet, trois types de corpus textuels[44] peuvent être simultanément analysés par la Station :

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

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

exploitation de différents corpus constitués dans un groupe de travail ayant des projets différents.

Sélection des outils

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

- les étiqueteurs morphosyntaxiques : statistique pour TreeTagger (SCHMID 1994) et à base de règles pour Brill[45] (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 forme 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 assorties d'une matrice 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) ;
- le racinisateur Lingua::Stem[46]: 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 pour 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.

Sélection de ressources terminologiques externes (prédéfinies)

Deux ressources externes sont actuellement prédéfinies dans la Station :

- TermSciences[47], portail terminologique multidisciplinaire développé par CNRS-INIST ;

[45] Avec le lexique et le fichier de règles fournis par l'ATILF-CNRS, de Nancy.
[46] http://search.cpan.org/~sdp/Lingua-Stem-Fr0.02/lib/Lingua/Stem/Fr.pm [04/12/2011].
[47] http://www.termsciences.fr/ [03/04/2014].

– IATE[48], 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[49]). 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.

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[50]) 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.

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 différents critères :

[48] http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load [03/04/2011].
[49] eurovoc.europa.eu [03/04/2011].
[50] Lexique de formes fléchies du français développé par ATILF, http://www.cnrtl.fr/lexiques/morphalou/LMF-Morphalou.php [03/04/2014].

- 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 ;
- 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)[51] dans le domaine analysé ;
- et la présence d'une ULC dans une ressource terminologique interne.

En voici le résumé dans le Tableau 3.

Paramètres de pondération	Explication
Nombre d'extracteurs ayant proposé une UL	
Base du PT par extracteur	Poids par extracteur ayant proposé une UL ; au carré pour 2 extracteurs, au cube pour 3 extracteurs. ***Exemple :*** *Pour la base du PT = 3, une UL attestée par 1 extracteur aura le poids de 3, par deux extracteurs 3^2= 9, par 3 extracteurs 3^3= 27*
Seuil de statut terminologique	Seuil à partir duquel une UL est considérée comme un terme
Attestation par une ressource terminologique	
Poids UL globale	Poids attribué à une UL lorsqu'elle est attestée par une ressource terminologique externe
Poids tête et expansion	Poids attribué à une UL lorsque sa Tête **et** son Expansion sont attestées par une ressource terminologique externe
Poids tête ou expansion	Poids attribué à une UL lorsque sa Tête **ou** son Expansion sont attestées par une ressource terminologique externe
Attestation dans un autre corpus	
UL présente dans le CS	Poids attribué à une UL présente dans le Corpus Support
UL présente dans le CC	Poids (négatif) attribué à une UL présente dans le Corpus Contrastif

Tableau 3 Paramètres de pondération du PT

(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.

[51] Estimée par l'analyste.

En voici le résumé dans le Tableau 4.

Paramètres de pondération	Explication
Attestation dans une ressource terminologique	
Poids UL globale	Poids (négatif) attribué à une UL lorsqu'elle est attestée dans une ressource terminologique externe
Matrice morphosyntaxique	
Poids verbe	Poids attribué à une UL dont la matrice morphosyntaxique est ou contient un verbe (Ver)
Poids participe	Poids attribué à une UL dont la matrice morphosyntaxique est ou contient un participe passé (ou présent) adjectival (Vppe ou Vppr)
Densité de la famille dérivationnelle	
Seuil d'UL dérivées	Seuil à partir duquel le poids d'UL dérivées est attribué
Poids d'UL dérivées	Poids attribué si le nombre d'UL dérivées distinctes de l'UL analysée dépasse le seuil *Exemple :* *À partir de 3 UL dérivées (seuil), on attribue le poids de 6 à l'UL analysée*
Densité de collocations	
Seuil d'UL collocatives	Seuil à partir duquel le poids d'UL collocatives est attribué
Poids d'UL collocatives	Poids attribué si le nombre d'UL collocatives distinctes de l'UL analysée dépasse le seuil *Exemple :* *À partir de 3 UL collocatives (seuil), on attribue le poids de 6 à l'UL analysée*
Extraction par Acabit	
Poids Acabit	Poids attribué à une UL extraite par Acabit

Tableau 4 Paramétrage de pondérations du PSL

(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.

En voici le résumé dans le Tableau 5.

Paramètres de pondération	Explication
Nombre d'extracteurs ayant proposé une UL	
Seuil de nombre d'extracteurs	Seuil à partir duquel le poids du nombre d'extracteurs est attribué
Poids du nombre d'extracteurs	Poids attribué à une UL lorsque le nombre d'extracteurs dépasse le seuil

Tableau 5 Paramétrage de pondérations du PUL

À 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 (cf. Figure 16).

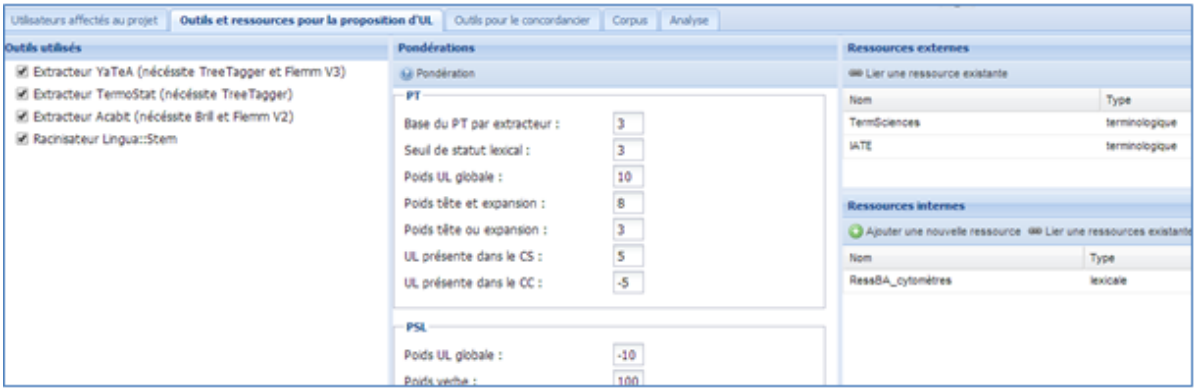


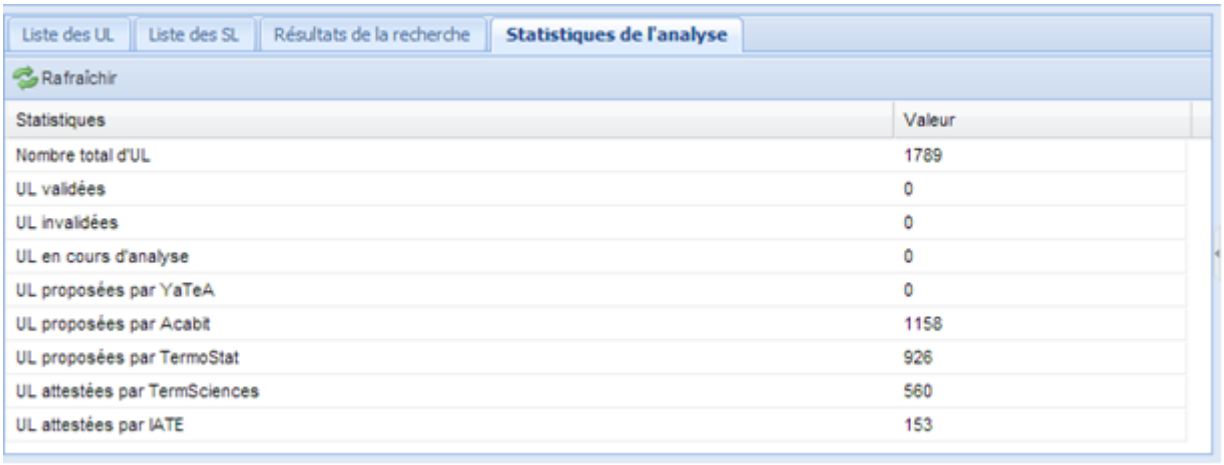
Figure 16 Station Sensunique, paramétrages des pondérations

L'algorithme définissant la configuration par défaut du calcul de trois pondérations se trouve dans l'Annexe 9.2.

3.4.3.3.2 Module d'Analyse automatique

Ce module a deux fonctions. La première fonction consiste à 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).

La Station affiche les **résultats quantitatifs** de l'analyse, à savoir le nombre total d'UL, le nombre d'UL par le statut de validation, ainsi que le nombre d'UL proposées par chaque extracteur ou attestées dans une ressource externe (Figure 17).



Statistiques	Valeur
Nombre total d'UL	1789
UL validées	0
UL invalidées	0
UL en cours d'analyse	0
UL proposées par YaTeA	0
UL proposées par Acabit	1158
UL proposées par TermoStat	926
UL attestées par TermSciences	560
UL attestées par IATE	153

Figure 17 Station Sensunique : Statistiques de l'analyse

Les résultats sont affichés de manière dynamique, c'est-à-dire qu'après chaque modification manuelle (ajout, suppression, validation, invalidation d'une UL), les statistiques sont mises à jour.

Les **informations qualitatives** issues de l'analyse automatique sont, pour chaque UL :

- **Forme Canonique** : correspond à la forme d'UL trouvée en corpus, priorisée dans l'ordre suivant : TermoStat, YaTea, Acabit. Normalement, la Forme Canonique devrait être la forme la plus simple d'une UL, utilisée, par exemple, comme entrée dans les dictionnaires. Cependant, aucun extracteur ne fournit de Forme Canonique définie de cette façon; d'où le choix d'utiliser comme Forme Canonique la forme trouvée en corpus par les extracteurs.
- **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 ;
    - 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* ; les variantes sont divisées en plusieurs types :

    - **Forme Abrégée** : il s'agit des abréviations (acronymes ou sigles) d'une UL analysée ;

Exemples
Formes abrégées de l'UL anticorps primaire : Ac Iaire ;
Formes abrégées de l'UL anticorps monoclonal : Ac Mo, AcM, ACMC ;

    - **Synonyme** : il s'agit d'UL répertoriées comme synonymes selon l'utilisateur ou selon une ressource attestée ;

Exemple
Synonyme de l'UL anticorps monoclonal : monoclonal ;

    - **Variante morphologique dérivationnelle** : une variante impliquant une dérivation entre un élément de 2 UL ;

Exemple
Variante morphologique dérivationnelle de l'UL marquage cellulaire : marquage de cellules (dérivation entre cellule/cellulaire)

- **Forme à variation syntaxique faible** : il s'agit de formes présentant des petits changements de structure, tels que l'insertion ou la variation (au sens changement) d'un mot grammatical ;

Exemple
Forme à variation syntaxique faible de l'UL cytomètre de flux : cytomètre en flux

- **Autre Variante** : il s'agit des autres types de variantes, non-recensés dans les types précédents, par exemple des variantes (correctes) orthographiques comme dans *anévrisme/ anévrysme*.

- **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* [52];
  - **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).

Les résultats de l'analyse automatique sont affichés dans l'Interface de travail de la Station Sensunique (Figure 18), divisée en 4 espaces :

Espace 1 : *Visualisation de la liste d'UL/SL*

- Permet de visualiser les résultats d'analyse automatique sous la forme d'une liste d'UL/ SL avec leurs champs associés ;
- Permet de gérer les relations entre les différentes UL ;
- Permet de visualiser, en corpus, les résultats de la recherche ;
- Permet de visualiser les résultats de l'analyse.

Espace 2 : *Filtres sur la liste d'UL/SL*

Espace 3 : *Visualisation d'UL en contexte*

- Permet de visualiser les UL sélectionnées en corpus;

[52] 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.

- Permet de visualiser les résultats de la recherche en corpus.

Espace 4 : *Visualisation de fiches lexicales d'UL/SL*

- Permet de visualiser et modifier les fiches d'UL/ SL.

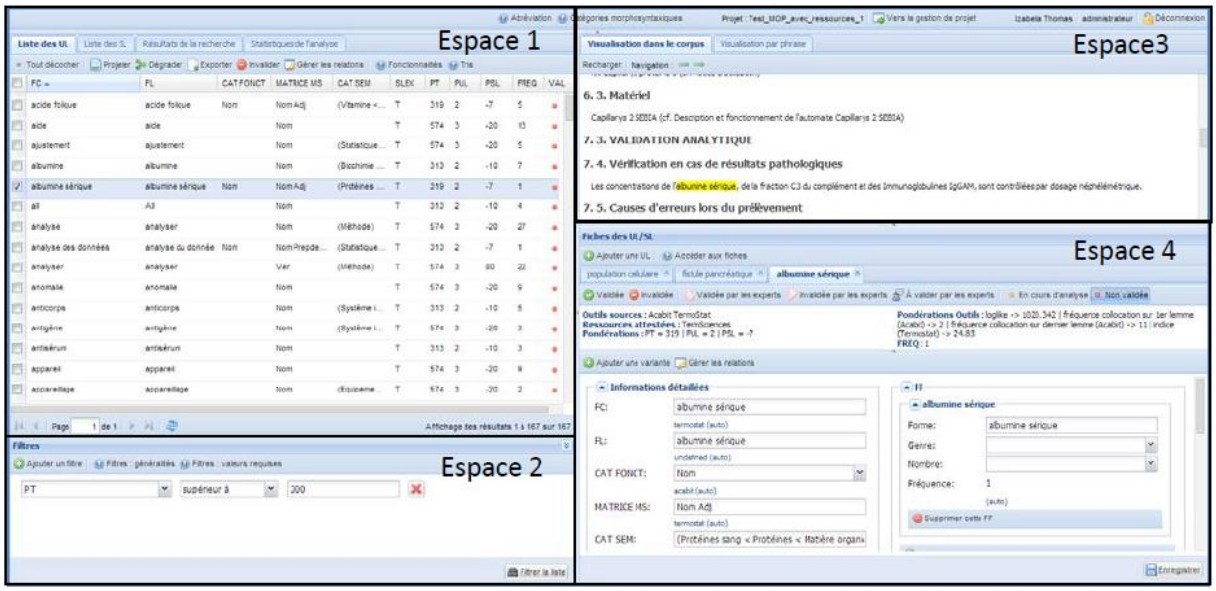


Figure 18 Station Sensunique : Interface de travail

Un cinquième espace *Recherche/filtres* peut être ouvert dans l'interface de travail afin d'effectuer des recherches sur le Corpus d'Analyse à l'aide d'un concordancier évolué (cf. Figure 19).

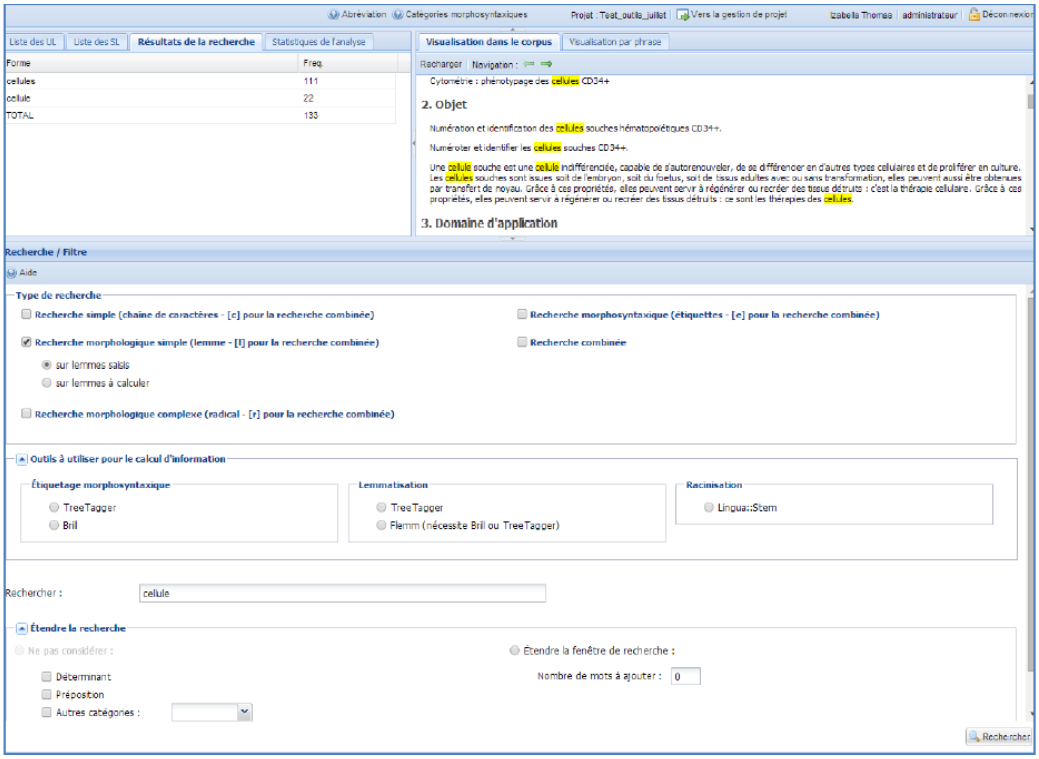


Figure 19 Station Sensunique : Concordancier évolué

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.

3.4.3.3.3 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 (cf. Figure 20) ;

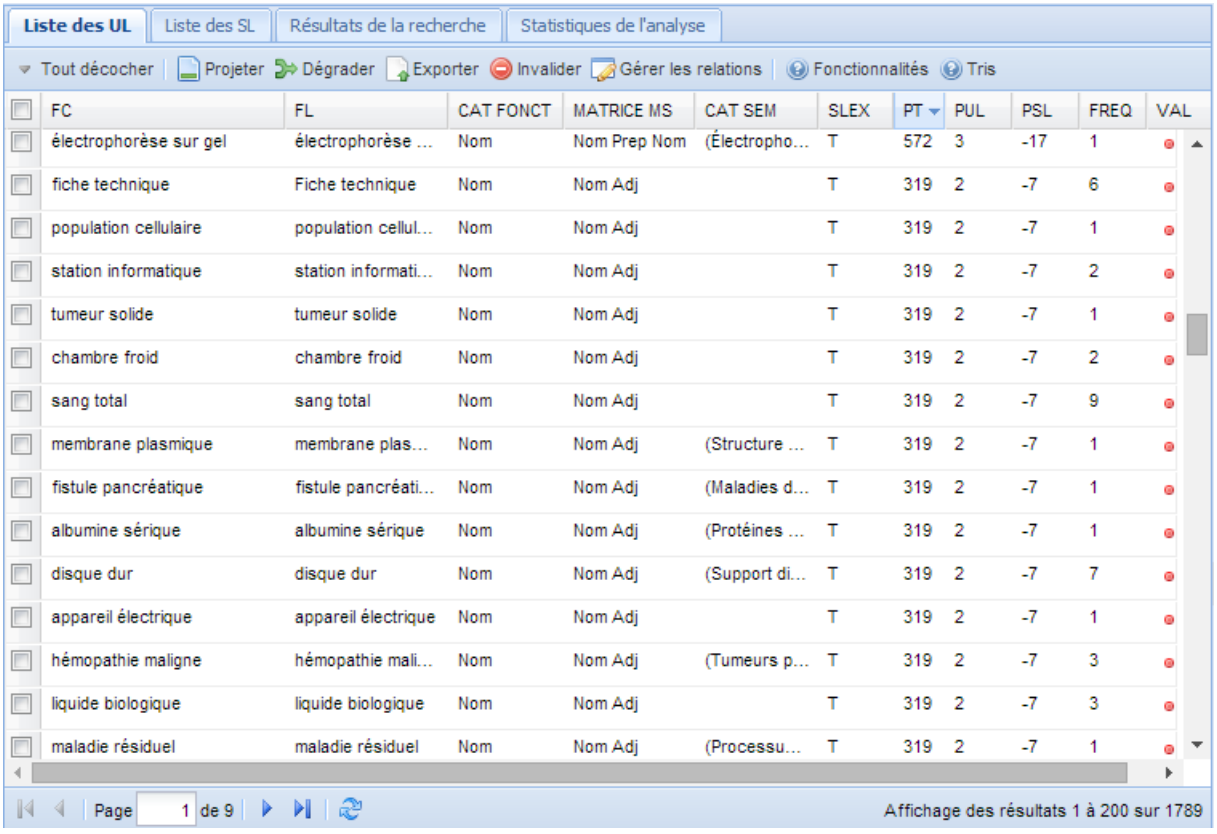


Figure 20 Station Sensunique : Liste des UL

– fiche lexicale de chaque UL/SL détaillant toutes les informations relatives à l'UL/SL analysée (cf. Figure 21) ;

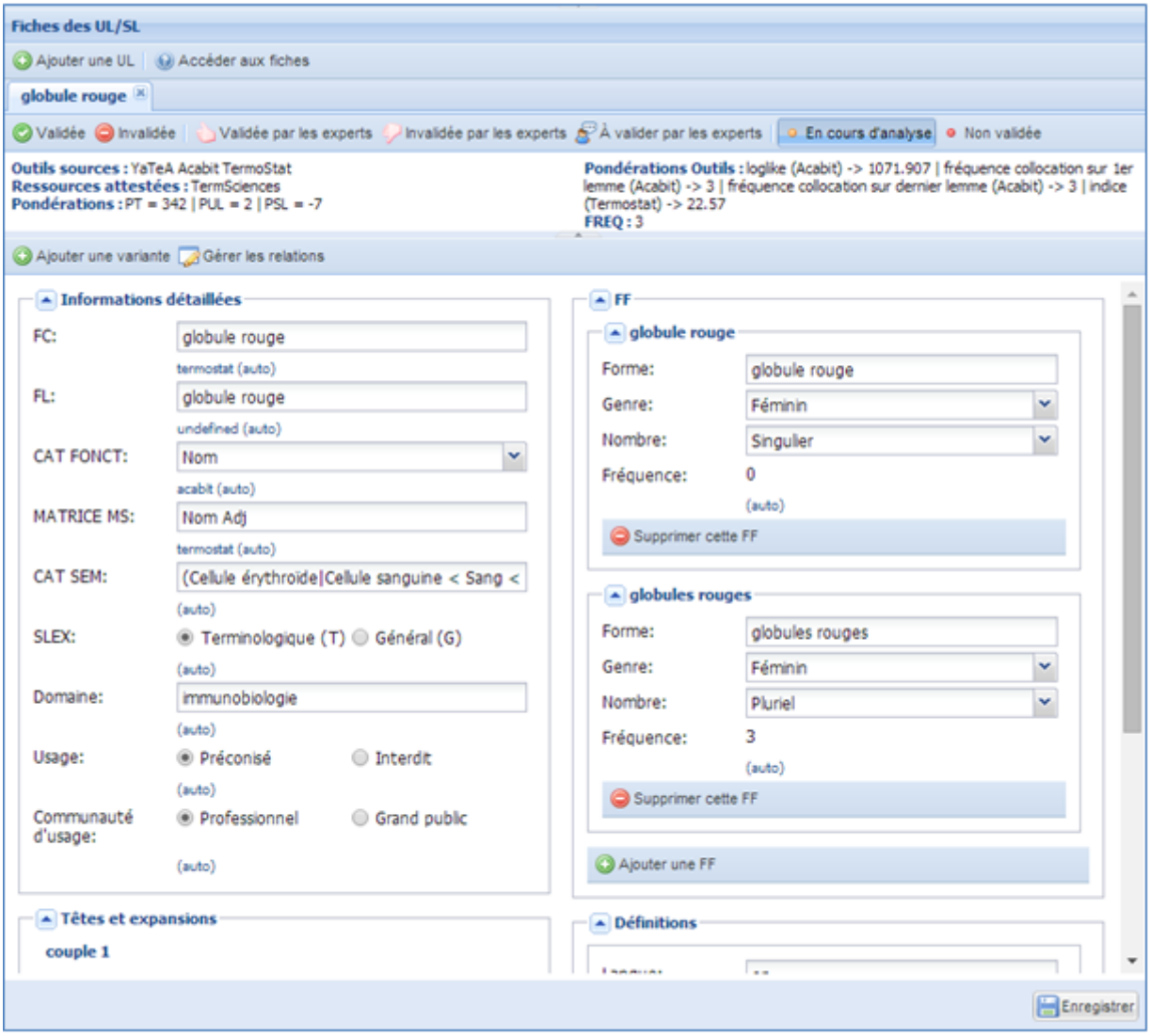


Figure 21 Station Sensunique : Fiche lexicale d'une UL

– 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) (cf. Figure 22).

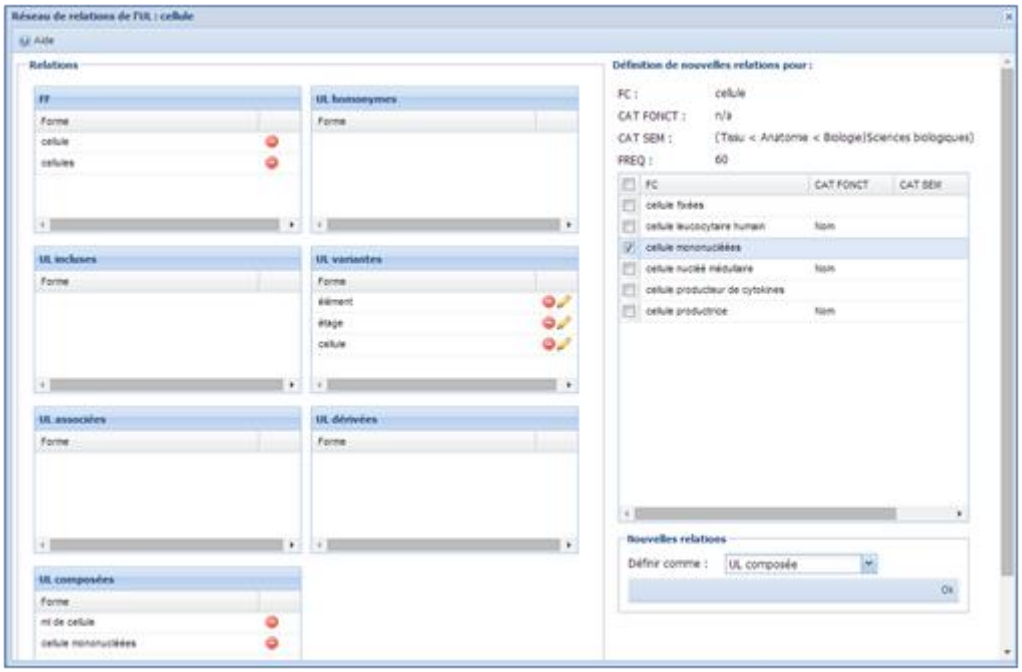


Figure 22 Station Sensunique : Fiche de relation d'une UL

L'analyste peut ajouter, modifier, compléter, valider ou supprimer toute ULC ou information. 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) (Figure 23) ;

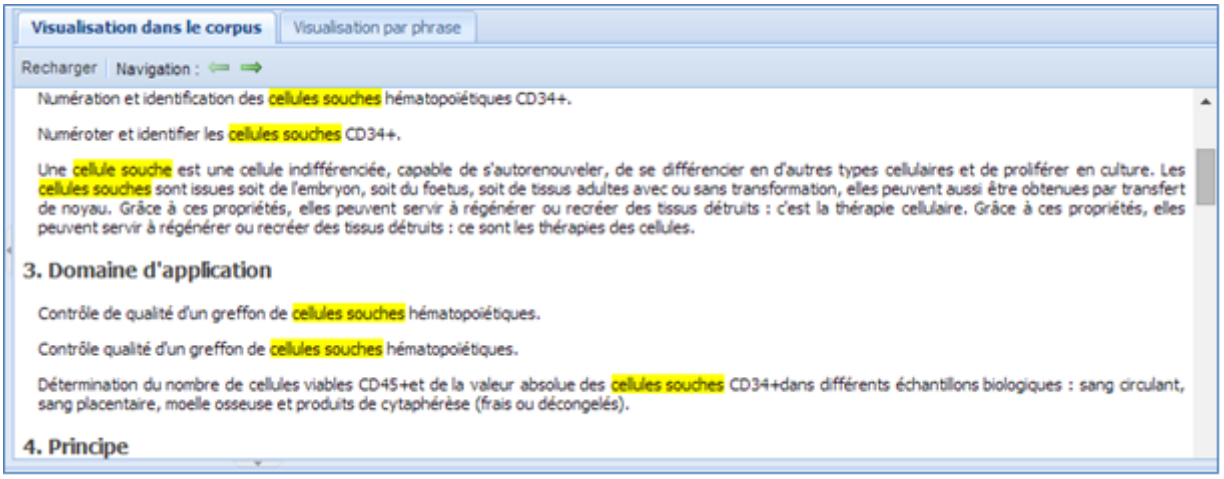


Figure 23 Station Sensunique : Visualisation des UL

- **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[53] :
    - (a) simple : sur une chaîne de caractères ;
    - (b) morphologique simple : sur un (ou une suite de) lemme(s) permettant d'identifier toutes les 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 recherche 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 (Figure 24).

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

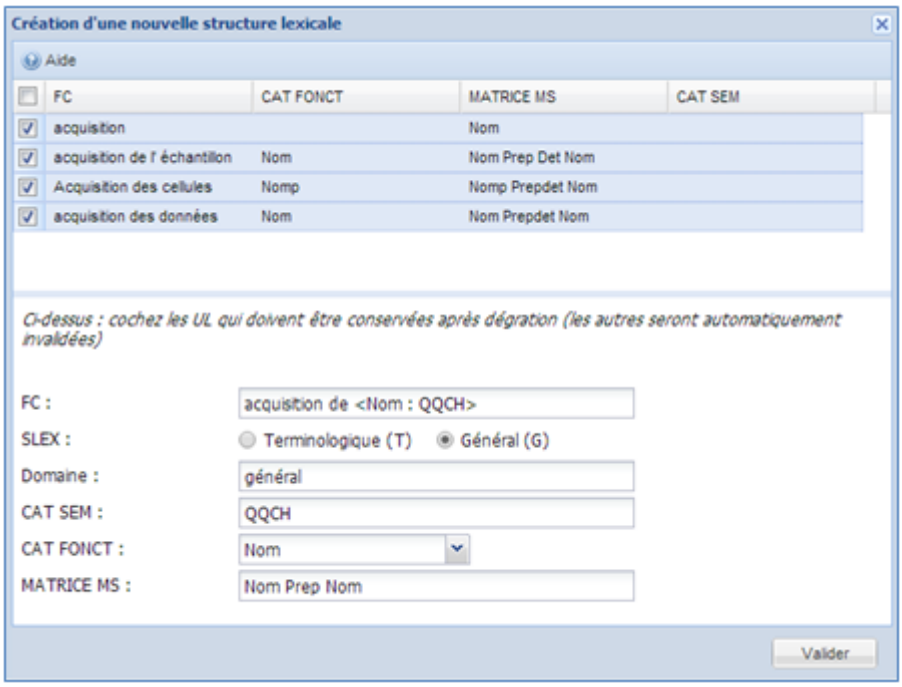


Figure 24 Station Sensunique : Création d'une SL

L'objectif du processus de validation est d'accepter ou de refuser les propositions d'UL/ SL issues de l'analyse automatique ou manuelle. 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, même s'il n'est pas obligatoire de passer par toutes les étapes de validation (Figure 25).

Figure 25 Station Sensunique : Validation d'une UL/SL

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 ;
- assurer l'interopérabilité des données ;
- 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[54]. 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 (Figure 26) :

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

Exemples
- dictionnaire d'UL contenant seulement : Forme canonique, Définition et Variantes

ou
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 »

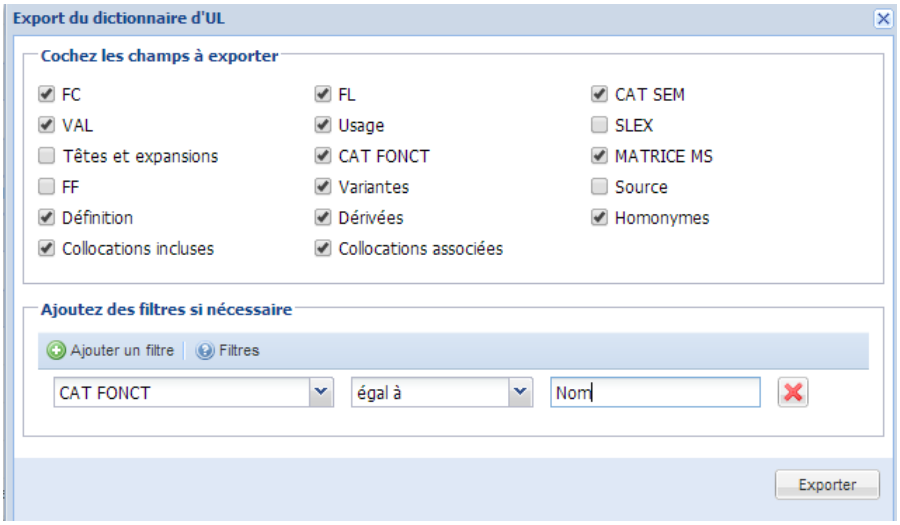


Figure 26 Station Sensunique : Export du dictionnaire d'UL

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

3.4.4 Conclusion

Conçue dans l'objectif d'optimiser (en termes de qualité et de coût) l'acquisition du lexique d'une LC, 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

[54] Voir un exemple de format de l'export en Annexe 9.3.

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 ressources terminologiques 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 la construction, sur le même modèle d'une station similaire pour d'autres langues et notamment pour l'anglais[55].

Pour le moment, bien que la Station ait été conçue pour faciliter l'acquisition du lexique contrôlé, nous l'utilisons comme un extracteur de termes dans nos travaux concernant les lexiques spécialisés. Le processus d'acquisition du lexique d'une LC est très proche de l'acquisition de ressources terminologiques tel que décrit par BOURIGAULT (2003). L'acquisition de dictionnaires, glossaires, lexiques, thesaurus à partir de corpus doit répondre à une double contrainte de pertinence, vis-à-vis du corpus et de l'application visée, (aide à la traduction, extraction d'information, indexation, etc.).

La Station répondant à ces contraintes, son champ d'application initial (lexique d'une LC) peut être élargi à l'ensemble des ressources terminologiques. 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

[55] C'est le projet du mémoire de Master de Steven DUGOIS, commencé en 2015 sous ma direction.

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

La construction d'une LC et, par conséquent, celle du lexique d'une LC correspondent à un besoin spécifique lié à la rédaction technique dans des domaines exigeant une haute sécurité. Dans ce contexte, le haut niveau de contrôle constitue un avantage, qui n'est cependant envisageable que dans le cadre d'une organisation puisqu'il se paye en temps de travail de conception d'une LC. Il existe d'autres contextes d'écriture en langue spécialisée, pour lesquels un tel investissement n'est pas nécessaire. Il s'agit de la rédaction de textes scientifiques.

4. REDACTION DE TEXTES SCIENTIFIQUES ET LEXIQUE SPECIALISE CONTEXTUALISE

Une très grosse partie de l'activité scientifique est consacrée non pas aux manipulations mais à la mise en forme écrite des résultats : un chercheur dépense la moitié de son temps à écrire […] Ce travail d'écriture, pendant lequel il est assis à son bureau, joue un rôle tout à fait déterminant dans la construction de la science.

Jacobi, 1993

L'intérêt pour l'étude des textes scientifiques n'est pas nouveau mais connait aujourd'hui un essor considérable. Ceci est, entre autres, lié à la disponibilité des données et à la construction de corpus qui permettent d'appréhender les écrits scientifiques sous différents angles (RINCK 2010). Celui qui nous intéresse particulièrement peut être à la fois situé dans une perspective didactique - former à la communication en langue étrangère de spécialité et à l'écrit académique, et applicative – aider les chercheurs à rédiger les articles scientifiques dans leurs langues spécialisées.

De quels types d'information avons-nous besoin pour rédiger correctement les articles scientifiques, surtout dans une langue étrangère ? Quelle forme devrait prendre un outil d'aide à la rédaction scientifique ? Quels sont les éléments de description nécessaires et suffisants pour améliorer la qualité des écrits ? Est-il utile de donner à l'utilisateur toutes les informations qu'on possède sur une lexie spécialisée ? Quelle quantité d'information est raisonnable? Comment ne pas tomber dans le défaut inverse des dictionnaires spécialisés d'aujourd'hui, à savoir ne pas noyer les utilisateurs sous un 'trop-plein' d'informations ? Aussi, quelle forme donner à l'information pour qu'elle soit facilement retenue par un utilisateur ?

Nous nous sommes posé ces questions à travers deux projets concernant les écrits scientifiques, que nous allons décrire dans la suite de ce chapitre. Le premier considère la conception d'un outil d'aide à la rédaction scientifique, destinés aux chercheurs 'en exercice', déjà rédigeant (&4.1). Le second a pour objectif de concevoir une plateforme d'aide à l'apprentissage du lexique spécialisé du niveau académique : il est plutôt destiné aux étudiants et apprenants d'une langue de spécialité (&4.2).

Selon CABRE (1998 : 133) le lexique est « le niveau le plus important » dans les écrits basés sur les langues spécialisées. RONDEAU (1983) surenchérit sur le fait que les langues spécialisées se caractérisent fondamentalement par leur lexique. Il existe énormément de travaux sur le lexique d'une langue de spécialité, et plus particulièrement sur le langage des écrits académiques. Ce langage peut être approché de différents points de vue, par exemple comme un genre textuel dans les travaux de POUDAT (2009, 2011), ou en analysant la structure du discours scientifique (BERTIN et al. 2015), ou encore en étudiant le positionnement de l'auteur[56].

[56] Voir les publications autours du projet Scientext regroupées dans TUTIN et GROSSMANN (2014).

Du point de vue du lexique, le langage scientifique se caractérise par la coexistence d'au moins trois types de lexique (BAKER 1988) :

- le lexique général qui correspond au vocabulaire de la vie de tous les jours ;
- le lexique terminologique ou spécialisé qui est spécifique à un domaine ;
- le lexique transdisciplinaire qui est commun à toutes les disciplines scientifiques.

Les lexiques terminologiques sont recensés dans les dictionnaires spécialisés et les banques terminologiques, papier et en ligne. Cependant, plusieurs remarques peuvent être faites par rapport à ces types de ressources. Premièrement, elles sont loin d'inventorier toutes les lexies spécialisées utilisées dans les écrits scientifiques d'un domaine (THOMAS et ATANASSOVA 2015). De plus, elles ne recensent pas beaucoup d'informations supplémentaires pouvant être utiles aux rédacteurs ; mais ce n'est peut-être pas non plus leur fonction.

Nous nous avançons sur nos résultats en affirmant que les outils que nous concevons cherchent à décrire les lexiques spécialisés **contextualisés**, quel que soit le lexique que nous construisons (lexique d'un domaine, lexique trans-biomédical etc.). Les termes sont caractérisés par « une syntagmatique restreinte (cooccurrences et commutations dans les limites d'un domaine spécialisé) » (LERAT 1995 : 52), il est donc possible d'énumérer les contextes significatifs d'une lexie spécialisée. Les travaux sur les collocations en langues spécialisées et sur les avantages que la description des collocations spécialisées peut apporter aux rédacteurs des articles témoignent de l'importance de ce sujet (TUTIN 2002, VOLANSCHI et KÜBLER 2010 ; KÜBLER et PECMAN 2012 ; L'HOMME 2013). Le choix de contextualisation du lexique spécialisé est aussi inspiré par des travaux concernant le lexique scientifique transdisciplinaire, qui donnent autant d'importance à l'étude du lexique qu'à l'analyse et le recensement de la phraséologie scientifique (TUTIN 2007). Nous avons aussi exploré du côté des études sur le recensement des cooccurrents dans les textes et sur la description des propriétés collocatives des unités lexicales dans les dictionnaires d'apprentissage, surtout dans les dictionnaires d'un type un peu nouveau comme DiCoInfo (L'HOMME 2008) ou le DAFA, Le Dictionnaire d'Apprentissage du Français des Affaires, développé à l'Université de Leuven en Belgique (BINON et al. 1992). Puisque nos travaux ont une visée applicative, l'ensemble de ces informations est traité en tenant compte des besoins des utilisateurs et des exigences des systèmes informatisés à mettre en place.

4.1 Système d'Aide à la rédaction scientifique (SARS) dans le domaine biomédical

L'attractivité et la qualité d'une recherche universitaire sont étroitement liées à la diffusion internationale de ses résultats. Aujourd'hui, la plupart des publications scientifiques se font en anglais, surtout en ce qui concerne les sciences dites 'dures', dont le biomédical fait partie. Cependant, la barrière linguistique constitue un obstacle important devant la publication en langue anglaise pour de nombreux chercheurs francophones. Des travaux de recherche ont été consacrés à la spécificité des textes scientifiques, modelés sur des standards anglo-saxons, et des manuels de rédaction scientifique sont proposés aux utilisateurs (cf. &

4.1.1). Les outils informatisés sont pour la plupart destinés aux traducteurs et consistent en des ressources généralistes, non-centrées sur la traduction scientifique : traducteurs automatiques, mémoires de traduction, bases de données terminologiques, les dictionnaires/glossaires électroniques mono/multilingues. Il existe aussi des outils informatisés plus spécifiques (cf. &4.1.1), mais ils sont totalement inconnus du public biomédical rédigeant (cf. & 4.1.2).

Pour décider du type de logiciel à développer, nous nous sommes appuyés sur les acteurs du terrain : les experts en rédaction médicale ainsi que les médecins, chercheurs et traducteurs affiliés au CHRU de Besançon et à la faculté de médecine de l'Université de Franche-Comté. En nous appuyant sur un état de l'art, nous avons mis en réflexion plusieurs questions, soit via une enquête en ligne, soit lors des réunions en présentiel.

Les choix principaux que nous avons discutés ont porté sur :

- Approche :
  - Monolingue ou français-anglais ;
- « Sous-ensemble » de la langue à traiter :
  - Langue scientifique transdisciplinaire, langue scientifique médicale, terminologie médicale ;
- Type de logiciel :
  - Edition (aide lors de la rédaction) ou révision.

4.1.1 État de l'art des aides (informatisées ou non) à la rédaction scientifique

Le domaine de la rédaction scientifique et, plus particulièrement, celui de la rédaction scientifique dans le biomédical est particulièrement bien décrit et documenté. La première source d'information pour les rédacteurs de textes scientifiques ce sont les manuels (pour n'en donner que quelques exemples : HUGUIER et al. 2003, MOSELEY 1991, SALMI et SALAMON 2012, etc.), des articles (ECARNOT et al.2015) et des recommandations d'éditeurs (The AMA Manual of Style ; The CSE Manual for Authors, Editors, and Publishers; The Chicago Manual of Style Online etc.) que l'on trouve facilement en format papier ou sur le Web. Il existe aussi des cours et des tutoriels adressés aux personnes qui souhaitent s'exercer en rédaction scientifique.

Dans ce type de publications, on trouve des recommandations concernant la structure et le contenu de chaque partie d'un article scientifique et des conseils directes concernant différents niveaux de la langue. Ces recommandations constituent des normes d'écriture adoptées par la communauté scientifique concernée.

Des conseils spécifiques sont donnés par rapport à chaque partie d'un article de type IMRaD ('*Introduction, Methods, Results and Discussion'*), qui est un format d'article quasi obligatoire pour les articles de recherche en biomédical.

Exemple
Règles concernant le **titre** :

- il doit être court (10 à 15 mots) ;
- les mots informatifs doivent être placés au début d'un titre, dans la "position forte" qui retient attention ;
- il ne faut pas mettre de mots "subjectifs" (*espoir*), d'abréviations, ni de références.
- il ne faut pas mettre dans un titre des mots qui ne sont pas informatifs, par exemple :
    - ✓ A propos de...
    - ✓ Considérations sur ...
    - ✓ Contribution à l'étude...
    - ✓ Le problème de ...
    - ✓ Revue sur...
    - ✓ Place de...
    - ✓ Intérêt de…
    - ✓ Utilisation de ...
    - ✓ Studies on the nature

En ce qui concerne la langue, il existe quelques règles morphologiques, surtout pour la définition des temps verbaux, par exemple utiliser le passé pour citer un autre auteur ("*Stere et al. ont montré que...")* ou utiliser le présent pour introduire un fait communément admis ou prouvé dans la littérature scientifique («*Il existe des formes familiales de la polyarthrite rhumatoïde*"). Mais la plupart de règles concernent le lexique et les structures à utiliser ou à éviter dans un langage scientifique.

Dans les règles concernant le lexique, on trouve :

- les règles concernant la **précision** dans le choix des **mots** :
    - éliminer les adjectifs et les adverbes imprécis, tels que *very, quite, rather, fairly, relatively, several, about, much* etc.;
    - supprimer les adjectifs et les adverbes creux (*un examen **attentif** du tableau*..) ;
    - remplacer les adjectifs imprécis (***grosse** tumeur* : 3 cm ? 15 cm ? 27 cm ? ; *observation **récente*** : 1970 ? 1980 ? 1985 ? ; *un **certain** nombre* : 10 ? 23 ? 67 ?) ;
    - proscrire la synonymie et la variation[57] (*tumeur, cancer, adénocarcinome, néoplames* = même lésion) ;
    - être attentif à la 'fausse' synonymie : les mots ne sont souvent synonymes qu'en apparence (*un cas* n'est pas *un patient*, *un sujet* ou *un individu*) ;
    - utiliser toujours les mots les plus simples (*nutriment --> aliment* ; *parachever --> terminer*) ;
    - faire attention à des mots apparentés :

variable/varié
égal/ équivalent à
suivant/selon
grâce à/ à cause de
syndrome/ symptôme/ symptomatologie
alternative/éventualité
différence (écart exprimé en nombre absolu)/ différentiel (écart exprimé en %)

[57] " Ne vous laisser pas emporter par les effets de style de votre langue maternelle. En anglais, n'hésitez pas à répéter un mot spécifique dans la même phrase." « Le mot exact prime sur le style. Un synonyme mal adapté est pire qu'une répétition." (MOSELEY 1991) ; "... la logique de la rédaction scientifique implique d'utiliser le même mot pour designer la même chose" (HUGUIER et al. 2003).

- les règles concernant les **groupes de mots** :
    - utiliser de bons collocatifs (exemple de Hugier et al. (2003 : 31)) :

Incorrect	Correct
Couverture antibiotique	traitement antibiotique ; antibiothérapie
Syndrome hyperthermique	fièvre
Ces résultats font apparaitre…	Ces résultats montrent…
L'appendice est en position sous-hépatique	L'appendice est sous-hépatique…
L'expérience a limité l'émergence des complications…	L'expérience a réduit les complications…
L'examen révèle… (on ne peut révéler que ce qui est caché...)	L'examen montre…
Le malade présente une céphalée…	Le malade a une céphalée…

    - surveiller les redondances :

strictement normal
prévu d'avance
traitement antibiothérapique
masse tumorale
tube complètement plein, etc.

- les règles concernant l'utilisation des **abréviations** :
    - utiliser la liste d'abréviations internationale pour les unités de mesure ;
    - éviter les abréviations qui ne sont utilisées que 3 ou 4 fois dans le texte : l'accumulation des abréviations, même si elles ont été expliquées, finit par rendre la lecture difficile ;
    - annoncer toute abréviation (*Une lithiase de la voie biliaire principale (VBP) a été observée...*) ;
- les règles concernant le **vocabulaire biomédical spécifique** :
    - pour les noms de médicaments : utiliser la dénomination commune internationale; écrire sans lettre majuscule; ex. amoxicilline; si le nom commercial est utilisé, écrire avec une lettre majuscule suivi de ® (Clamoxyl®) ;
    - pour les noms de bactérie ou d'animal comportant deux noms latins : écrire en italique, première lettre du premier mot en majuscule ; ex. *Streptococcus viridans.*

Les règles concernant les structures préconisent de :

- éviter certains types de formulations ; par exemple éviter les formulations à la troisième personne qui alourdissent le texte (*one notices that..., one may see that…, it may be seen that...)* et les remplacer par les formules plus concises (*as shown in..., as indicated in.., as demonstrated in.., as presented in...*) *;*
- éviter le **passif de modestie**[58] et le remplacer par la voix active ;
- supprimer les **expressions creuses** :

Il parait utile/ opportun/ intéressent de remarquer / signaler que..

[58] "En rédaction scientifique, le passif de modestie expose aux mêmes ambigüités que le présent narratif. Lorsqu'un auteur écrit: "Dix malades ont été examinés par échographie", le lecteur ne sait pas si l'examen échographique a été réalisé par l'auteur de l'article, ou par quelqu'un d'autre." (HUGUIER et al. 2003 : 25).

Il va sans dire que...
Un certain nombre de points nous semblent mériter discussion...
We wish to call attention to the fact...

- éviter les **formulations elliptiques** :

Les principaux symptômes des 1006 cancers opérés --> Les principaux symptômes des 1006 malades opérés de cancers

- éviter l'utilisation des **adjectifs successifs** :

Le cirrhotique ascitique infecté → *qu'est-ce qui est infecté* : le cirrhotique *ou* l'ascite du cirrhotique ?
L'hypertendu artériel grave : *qu'est-ce qui est grave ?*

- bannir les '**expressions émotionnelles'** (HUGUIER et al. 2003 : 26) :

Nous avons déploré...
Par malheur, par chance...
Malheureusement, heureusement...
Nous avons eu la surprise de constater que...
Nous avons été étonnés...
Le malade a bénéficié/a subi → le malade a eu
Il est important de noter…
Intéressant, passionnant, décevant
Curieux, surprenant

- éviter des '**expressions lourdes**' :

de façon appréciable → beaucoup
l'ensemble de → tous
de manière à, de façon que, afin que → pour
il est indispensable → il faut
fournir une indication→ indiquer
en conséquence, pour ces différentes raisons → donc

- simplifier :

it is believed to have an essentiel role → it has an essential role
appears to play a role as (a central mediator) → is (a central mediator)
The diagnosis (..) was established based on → The diagnosis (..) was based on

Il est très facile de remarquer la ressemblance entre ces recommandations et les règles d'écriture d'une langue contrôlée, surtout lorsqu'on lit dans le manuel de Moseley (1990) : « *Émettre une bonne idée par phrase est l'une des meilleurs méthodes utilisées par les plumes scientifiques expérimentées* ». On y retrouve les objectifs de clarté, simplicité, lisibilité, et une certaine neutralité dans le positionnement des auteurs. Certes, il ne s'agit pas de contrôle strict, mais les recommandations peuvent être très poussées et précises (telles que *ne pas dépasser 50 mots par phrase ; l'introduction doit contenir maximum 3 paragraphes*, etc.). Certains conseils restent flous (*'Écrivez ce qu'il faut, ni plus, ni moins'*), mais ils s'adressent à des rédacteurs humains et comptent sur leur discernement. D'autres pourraient être formalisés et informatisés ('Ne pas mettre dans un titre des mots qui ne sont pas informatifs : *à propos de, considérations sur, place de, studies on the nature of…')*, et d'autres encore, non-informatisables, rappelés aux auteurs lors du processus de l'écriture, comme c'est déjà le cas dans le Compagnon LiSe (cf.&3.3.1).

Même s'il existe une énorme documentation concernant la rédaction biomédicale et plus généralement, la rédaction scientifique, nous n'avons trouvé que très peu de logiciels qui lui sont expressément dédiés. La plupart des outils que nous avons identifiés sont consacrés à la gestion de la communication scientifique (EasyChair, Manuscript Manager etc.), à la gestion des références bibliographiques (EndNote, Zotero etc.) ou à la recherche de références (MEDLINE/PubMed, Web of Science etc.). Nous avons donc fait un état de l'art très large, en nous inspirant des logiciels qui ne sont pas toujours dédiés à la rédaction scientifique biomédicale. Les logiciels que nous allons présenter peuvent être regroupés selon les fonctionnalités que l'on peut mettre en avant, concernant la rédaction scientifique :

- fonctionnalité d'édition (Amadeus ; COBWEB) ;
- fonctionnalité d'aide lors de la rédaction (base de données ARTES ; projet Scientext ; Linguee ; TradooIT) ;
- fonctionnalité de révision et vérification (SWAN).

Interfaces d'édition

AMADEUS est une suite logicielle, développée à l'université de Sao Paulo au Brésil à partir des années 1990, et composée de plusieurs outils qui ont pour objectif d'aider les locuteurs non-natifs à produire des articles scientifiques en anglais (ALUISIO et al. 2001 ; OLIVEIRA JR et al. 2006). Basé sur l'analyse de corpus, l'outil principal de la suite n'est pas tant un outil centré sur les problèmes langagiers, mais sur la difficulté de comprendre et produire la structure interne d'un discours scientifique. Par conséquent, l'outil propose une aide à la structuration du discours scientifique (Figure 27) sous forme de traits spécifiques à des stratégies rhétoriques identifiées dans des discours existants. L'utilisateur compose sa propre stratégie rhétorique, qui est ensuite évaluée par l'outil sur la base de cas déjà identifiés dans sa base de données.

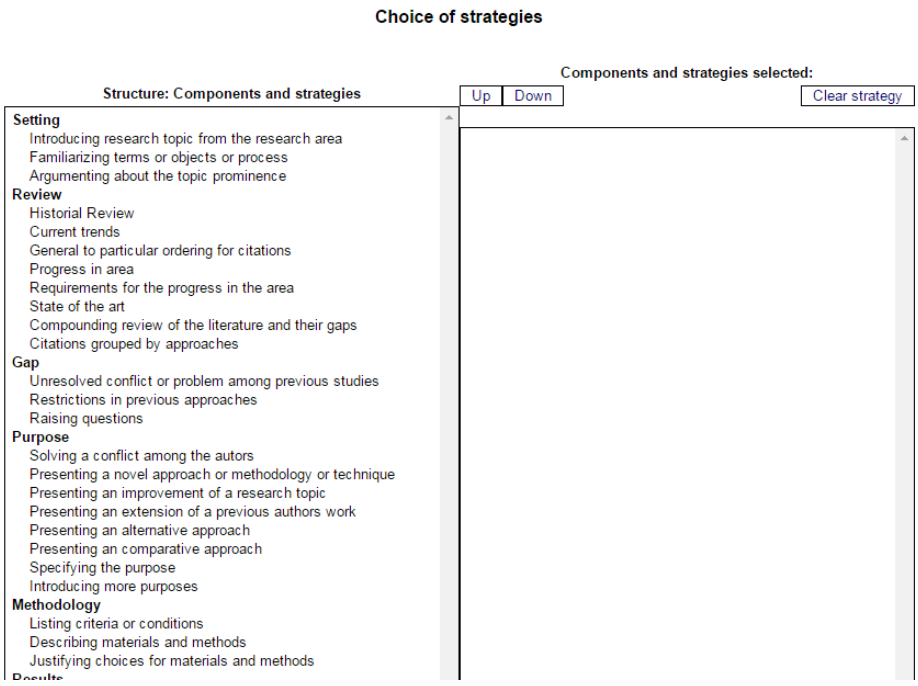
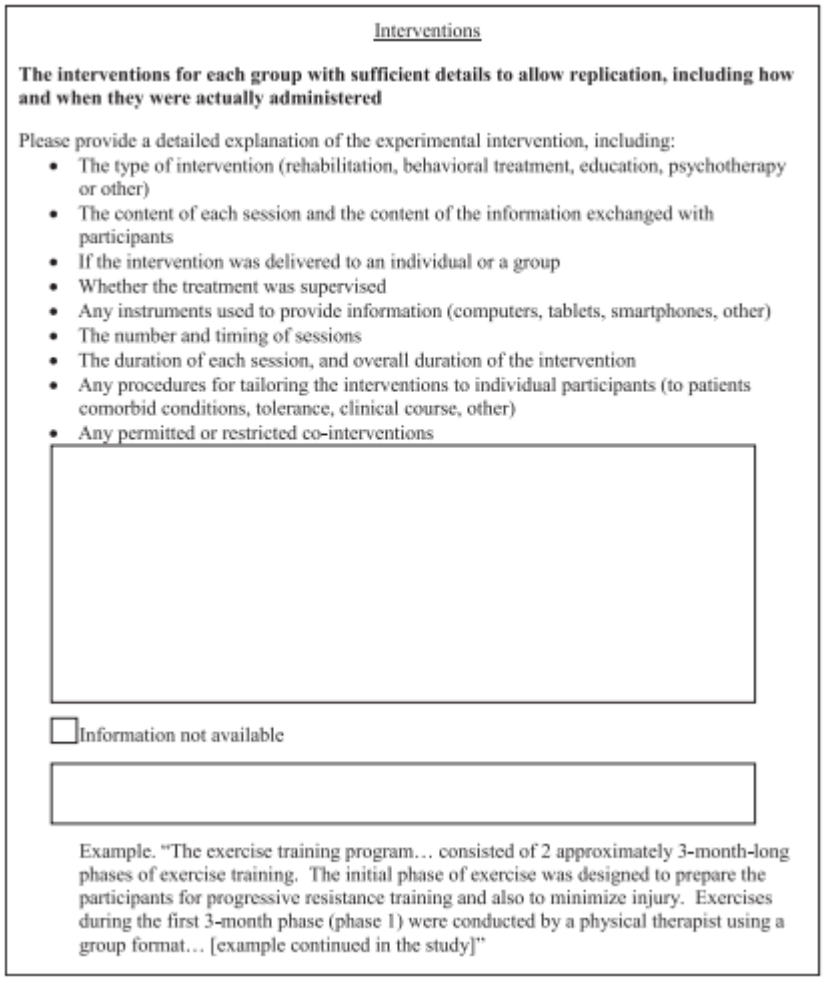


Figure 27 Amadeus : Choix de structures rhétoriques

Dans le domaine biomédical, BARNES et al. (2015) décrivent une interface très spécialisée, COBWEB, conçue pour aider les rédacteurs à faire des comptes rendus des essais contrôlés randomisés. Ce type d'écrits est hautement standardisé, puisque l'exactitude et la complétude des informations sont essentielles pour interpréter correctement les résultats de la recherche et leurs conséquences sur la pratique clinique. Par conséquent, les guides et les check-lists sur les informations requis dans ce type d'écrits sont proposés aux auteurs. Cependant, les auteurs ont du mal à respecter les règles, puisqu'elles sont nombreuses et noyées dans une masse de documentation. D'où l'idée d'une interface d'édition qui rappelle à l'auteur les informations qui doivent être fournies (Figure 28). En guidant les auteurs à travers les recommandations, le logiciel assure que les principales informations seront incluses dans les rapports.


Interventions

The interventions for each group with sufficient details to allow replication, including how and when they were actually administered

Please provide a detailed explanation of the experimental intervention, including:

- The type of intervention (rehabilitation, behavioral treatment, education, psychotherapy or other)
- The content of each session and the content of the information exchanged with participants
- If the intervention was delivered to an individual or a group
- Whether the treatment was supervised
- Any instruments used to provide information (computers, tablets, smartphones, other)
- The number and timing of sessions
- The duration of each session, and overall duration of the intervention
- Any procedures for tailoring the interventions to individual participants (to patients comorbid conditions, tolerance, clinical course, other)
- Any permitted or restricted co-interventions

☐ Information not available

Example. "The exercise training program… consisted of 2 approximately 3-month-long phases of exercise training. The initial phase of exercise was designed to prepare the participants for progressive resistance training and also to minimize injury. Exercises during the first 3-month phase (phase 1) were conducted by a physical therapist using a group format… [example continued in the study]"

Figure 28 COBWEB : Interface d'édition pour les essais randomisés contrôlés

Cette interface a été évaluée positivement par rapport à la complétude des rapports écrits à l'aide de l'interface en opposition aux rapports écrits sans son aide (BARNES et al. 2015). Cependant, même si très utile, cette interface est centrée sur les connaissances et non sur le langage : elle n'utilise pas de techniques de TAL et n'influence pas le langage utilisé lors de la rédaction.

Aide lors de la rédaction

L'analyse des textes scientifiques a fait l'objet de deux projets français ARTES (2007-2011) et Scientext (2007-2010), qui ont, entre autres, développé des outils d'aide à la rédaction scientifique.

La base de données ARTES (Aide à la Rédaction de TExtes Scientifiques), consultable en ligne[59] a été développée à l'Université Paris Diderot (PECMAN et KÜBLER 2011) pour répondre aux besoins de rédaction et de traduction en langues spécialisées. Elle propose deux types d'information via les interfaces associées :

1. ***Terminologie en contexte*** permet de rechercher des termes en plusieurs langues (Figure 29) et d'afficher les informations qui leur sont associées, entre autres : la définition, le contexte d'utilisation et la traduction ;

Figure 29 ARTES : Terminologie en contexte

2. ***Phraséologie discursive*** permet de faire une recherche sur les fonctions sémantico-discursives partagées par les diverses sciences, et qui donnent lieu à une analyse des collocations génériques servant à organiser le discours, annoncer les grandes lignes de l'étude, exprimer une temporalité, exprimer un point de vue, etc. L'utilisateur dispose d'une centaine de fonctions : par exemple, lorsqu'on choisit la fonction 'Présenter ses objectifs de recherche' (Figure 30), une centaine de collocations associées à cette fonction s'affichent dans une langue choisie par l'utilisateur (en français, en occurrence).

[59] https://artes.eila.univ-paris-diderot.fr/, accédé le 29/07/2016.

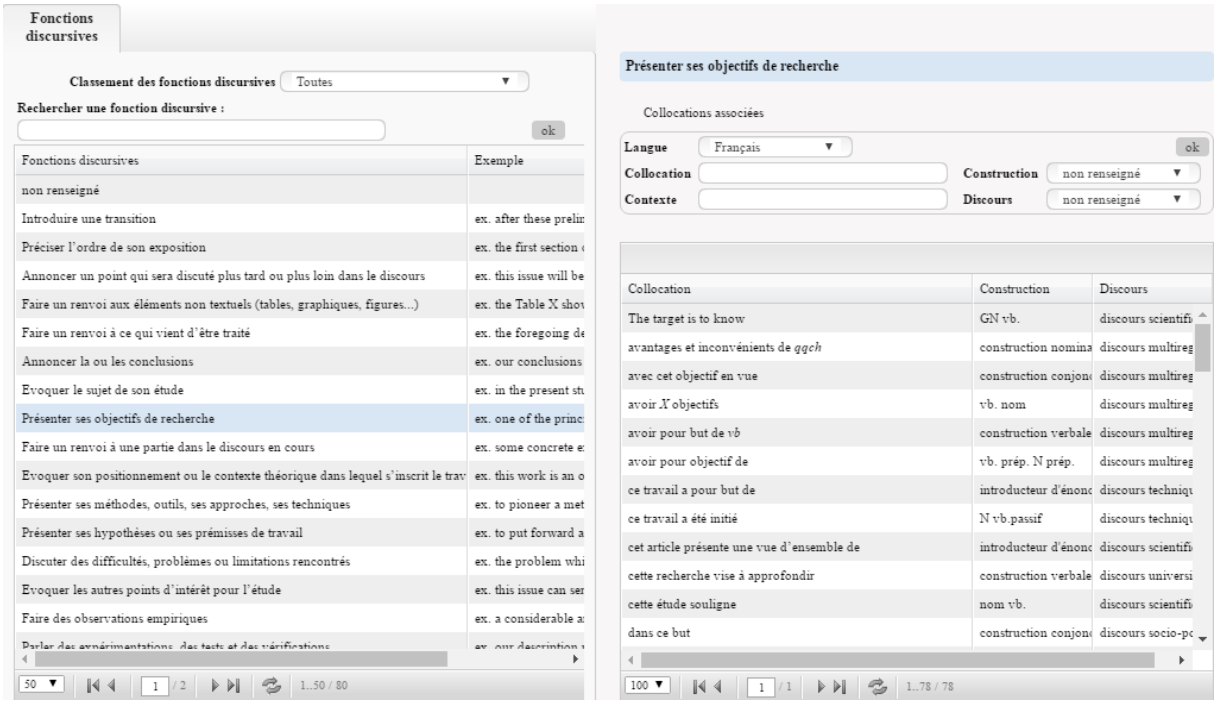


Figure 30 ARTES : Phraséologie discursive

Il est aussi possible de filtrer les collocations par le type de structures syntaxiques (cf. *construction verbale* sur la Figure 31) et type de discours (juridique, administratif, universitaire etc.). Sur la fenêtre de gauche on peut visualiser la collocation choisie et sa traduction sur des exemples tirés d'un corpus.

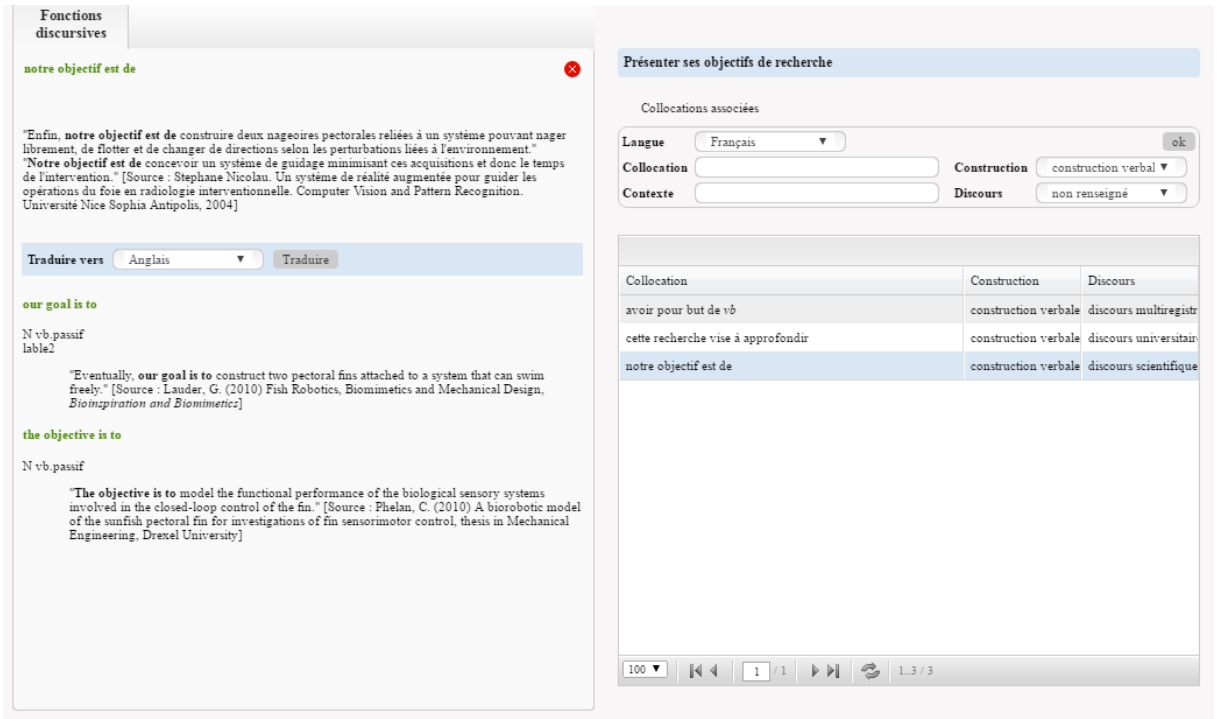


Figure 31 ARTES : Visualisation en corpus

Le projet ANR Scientext[60] met à la disposition des chercheurs un large corpus **d'écrits scientifiques** consultable en ligne, d'environ 4,8 millions de mots en français —la partie anglaise étant encore plus importante (33 millions de mots). Il comporte des écrits choisis selon les genres (thèse, articles, actes de colloque, écrits d'étudiants) et selon la discipline. L'objectif étant de créer un corpus multidisciplinaire, le corpus comprend les textes scientifiques dans plusieurs disciplines des sciences humaines, expérimentales et sciences pour l'ingénieur. Il est annoté au niveau de la structure des écrits et aux plans morphologique et syntaxique. L'objectif premier du projet Scientext a été d'étudier, en s'appuyant sur ce corpus, le positionnement et le raisonnement de l'auteur d'un écrit scientifique, à travers la phraséologie, les marques énonciatives et les marques syntaxiques liées à la causalité (TUTIN et GROSSMANN (eds) 2014). Il a aussi donné lieu à des études sur l'aspect didactique d'un tel corpus et son utilisation pour l'enseignement.

Le projet s'est doté d'une plateforme ScienQuest (FALAISE et al. 2012) qui vise à accompagner l'exploration linguistique du corpus par des non-spécialistes, qui sont définis comme des personnes ayant peu de connaissances en informatique. Par ailleurs, ils peuvent être « … *des TAListes en train de construire des grammaires, linguistes étudiant la distribution d'un phénomène linguistique, littéraires en pleine étude de style, apprenants d'une langue étrangère souhaitant vérifier l'usage d'un terme ou d'une tournure de phrase…* » (FALAISE et al. 2012 : 105). L'interface a été donc pensée en termes d'absence de technicité, de rapidité et facilité d'emploi et d'expressivité et progressivité.

ScienQuest permet de sélectionner le sous-corpus et les textes sur lesquels on veut l'interroger, ainsi que la partie des textes scientifiques que l'on veut exploiter (parmi Introduction, Développement, Conclusion et Résumé). Trois modes de recherche sont proposés à l'utilisateur :

- Recherche sémantique, qui s'appuie sur des grammaires locales prédéfinies autour d'une notion sémantique pertinente l'analyse de textes scientifiques, telle que l'expression d'une hypothèse, verbes de choix et d'intention, adjectifs d'évaluation etc.
- Recherche libre guidée, qui permet des recherches sur les mots et les relations syntaxiques qu'ils entretiennent. Par exemple, on peut rechercher des verbes qui ont comme objet direct le mot *cells* (Figure 32). Théoriquement, on peut ajouter autant de mots et de relations qu'on souhaite.

[60] http://scientext.msh-alpes.fr/scientext-site/spip.php?article1, accédé le 30/07/2016.

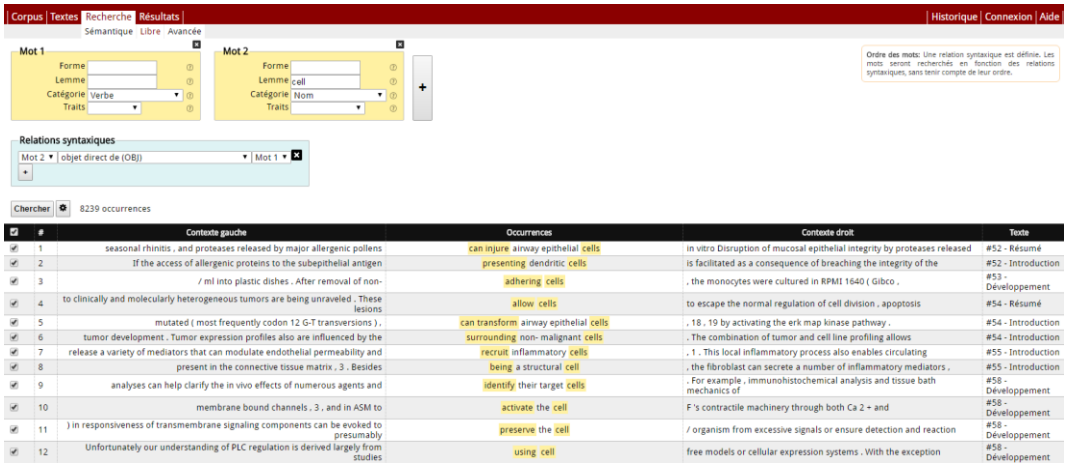


Figure 32 Scientext : Recherche libre guidée

- Recherche avancée, qui permet de créer directement une requête complexe, en suivant un langage de requête prédéfini. Ce mode est destiné aux utilisateurs spécialistes, linguistes familiarisés avec le TAL ou avec les traitements formels, informaticiens ayant des connaissances linguistiques.

Les résultats des recherches sont affichés dans un format classique de concordanciers, KWIC (Key-Word-in-Context).

Deux autres outils, non directement issus de la recherche académique, mais s'appuyant sur ses résultats sont intéressants à mentionner : Linguee et TradooIT.

Linguee[61] est un concordancier bilingue, qui fournit instantanément des traductions en contexte pour un mot ou un groupe de mots (Figure 33). Les résultats, sous forme de phrases alignées, proviennent d'une recherche sur un corpus de textes alignés que le logiciel compile sur le Web. L'outil ne garantit pas la qualité des résultats, mais laisse à l'utilisateur le choix de la traduction à partir des exemples fournis. Il dispose aussi d'un dictionnaire rédactionnel qui propose des traductions vérifiées, lorsqu'elles existent.

[61] http://www.linguee.fr/francais-anglais, accédé le 02/08/2016.

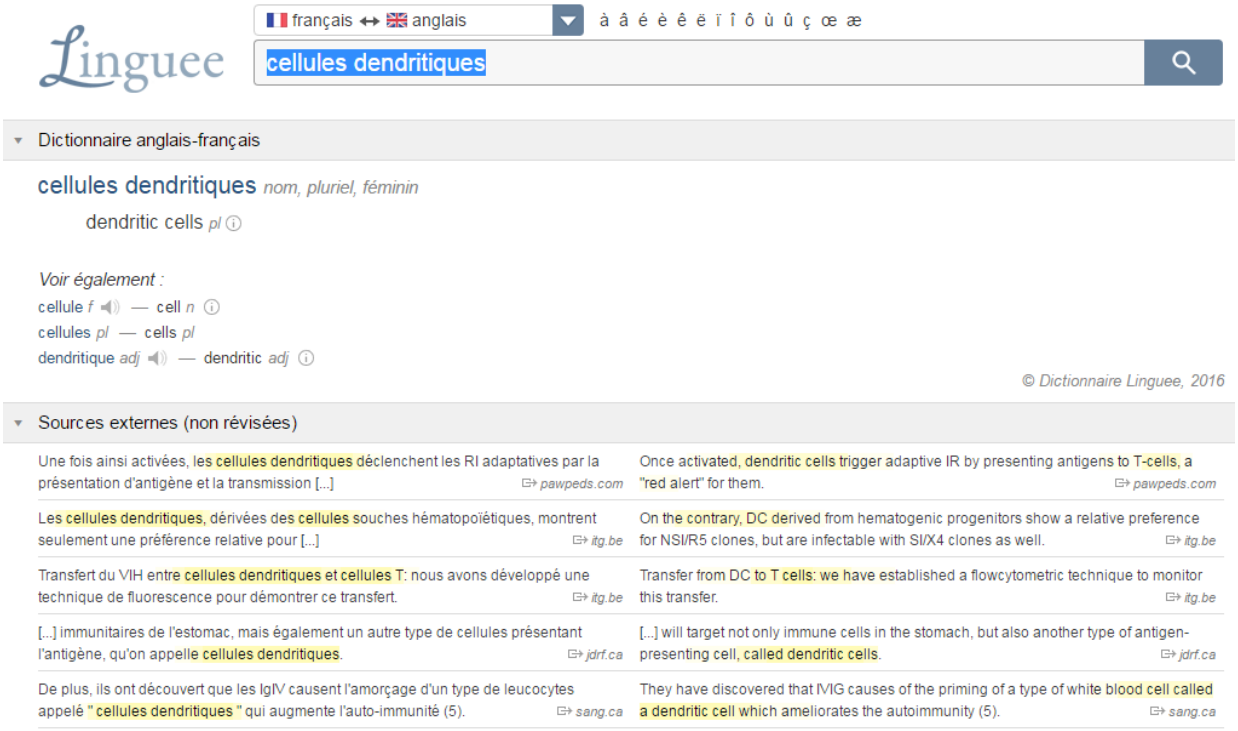


Figure 33 Linguee : Interface de recherche

TradooIT[62] est une suite d'outils de traduction assistée par ordinateur qui regroupe notamment une mémoire de traduction, une banque de terminologie et un concordancier bilingue (Figure 34). Par rapport à Linguee, TradooIT est plus orienté terminologie : le concordancier bilingue de TradooIT permet de rechercher des chaînes en contexte dans la mémoire de traduction et de trouver des termes dans les banques de terminologie (Termium[63], Onterm[64], GDT[65] etc.). Il présente ensuite les résultats sous forme de tableau, où il surligne les occurrences en langue de départ et leur équivalent en langue d'arrivée. Il présente aussi les statistiques des différentes formes et traductions trouvées et permet de filtrer les résultats en fonction de la forme, de la source, etc., pour préciser la recherche. Il offre également aux utilisateurs avancés la flexibilité de rechercher les radicaux des mots et d'inclure des variables dans leur expression de recherche. Enfin, le concordancier comporte des fonctions connexes, dont l'une propose d'autres recherches susceptibles d'être plus fructueuses en cas de résultats insuffisants, et l'autre suggère une traduction pour l'expression recherchée en cas de certitude suffisante.

[62] https://www.tradooit.com/index.php, accédé le 02/08/2016.
[63] Termium Plus, la banque de données terminologiques et linguistiques du gouvernement du Canada, http://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra, accédé le 02/08/2016.
[64] Onterm, source de terminologie officielle du gouvernement de l'Ontario, https://www.sdc.gov.on.ca/sites/mgcs-onterm/fr/Pages/default.aspx, accédé le 02/08/2016.
[65] Le grand dictionnaire terminologique (GDT), banque de fiches terminologiques rédigées par l'Office québécois de la langue française, http://gdt.oqlf.gouv.qc.ca/index.aspx, accédé le 02/08/2016.

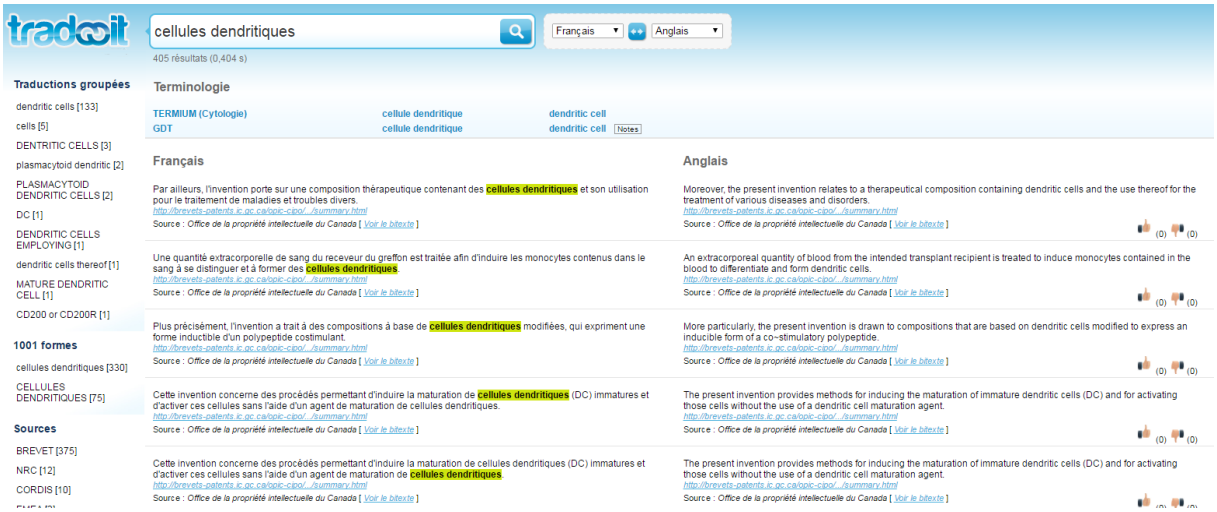


Figure 34 TradooIT : Interface de recherche

Linguee et TradooIT sont représentatifs d'un nouveau type de logiciels, développés suite à l'insuffisance (en termes de services rendus) des logiciels de traduction automatique et des dictionnaires classiques[66]. Ils laissent plus de liberté à l'utilisateur et proposent une stratégie nouvelle d'aide à la traduction, basée sur l'exemple. C'est cette stratégie, accompagnée d'une analyse linguistique des données, qui nous intéresse dans le projet de construction d'un outil d'aide à la rédaction.

Révision

Le logiciel SWAN (Scientific Writing AssistaNt), conçu par J.L. LEBRUN (2011) est un projet finlandais de logiciel d'identification et de correction d'éventuels problèmes d'écriture liés à la rédaction de textes scientifiques (KINNUNEN et all. 2012)[67]. Il se présente comme une interface d'édition basée sur les règles, qui parcourt les parties clés d'un article scientifique, à savoir le titre, l'abstract, l'introduction et la conclusion et donne à l'auteur un retour sur leur contenu. Les différentes mesures sont développées en fonction de la partie de l'article analysée. Par exemple, l'introduction est évaluée par rapport (TURUNEN 2013) (Figures 35 et 36) :

- à sa longueur (elle doit être concise, mais non incompréhensible) ;
- aux idées qui sont développées au début et à la fin de l'introduction (par exemple, il ne faut pas décrire la méthodologie à la fin de l'introduction, puisque le paragraphe sur la méthodologie suit l'introduction) ;
- à sa complétude (est-ce que l'introduction répond à sa fonction, c'est-à-dire qu'elle répond aux questions de pourquoi de la recherche) ;
- à sa précision (est-ce que les auteurs n'utilisent pas de mots imprécis (ex :. *generally, commonly, can, may*), de mots de jugement (*'judgement words'*, ex :.

[66] Il en existe d'autres, moins connus, par exemple Wordscope (http://www.wordscope.com/) ou Reverso Context (http://context.reverso.net/traduction).

[67] Il existe d'autres logiciels de vérification et de correction, mais qui ne sont pas directement liés à la rédaction scientifique ; voir par exemple la plateforme Accept (http://accept-portal.unige.ch/) pour la pré- et post-édition, ainsi qu'un logiciel industriel Captilo (http://www.prolipsia.com/captilo/) destiné à améliorer la qualité de textes techniques.

fail, unreliable, suffer), de surévaluation (‘*overstatements’*, ex :. *absolutely, certainly, acute*), etc.

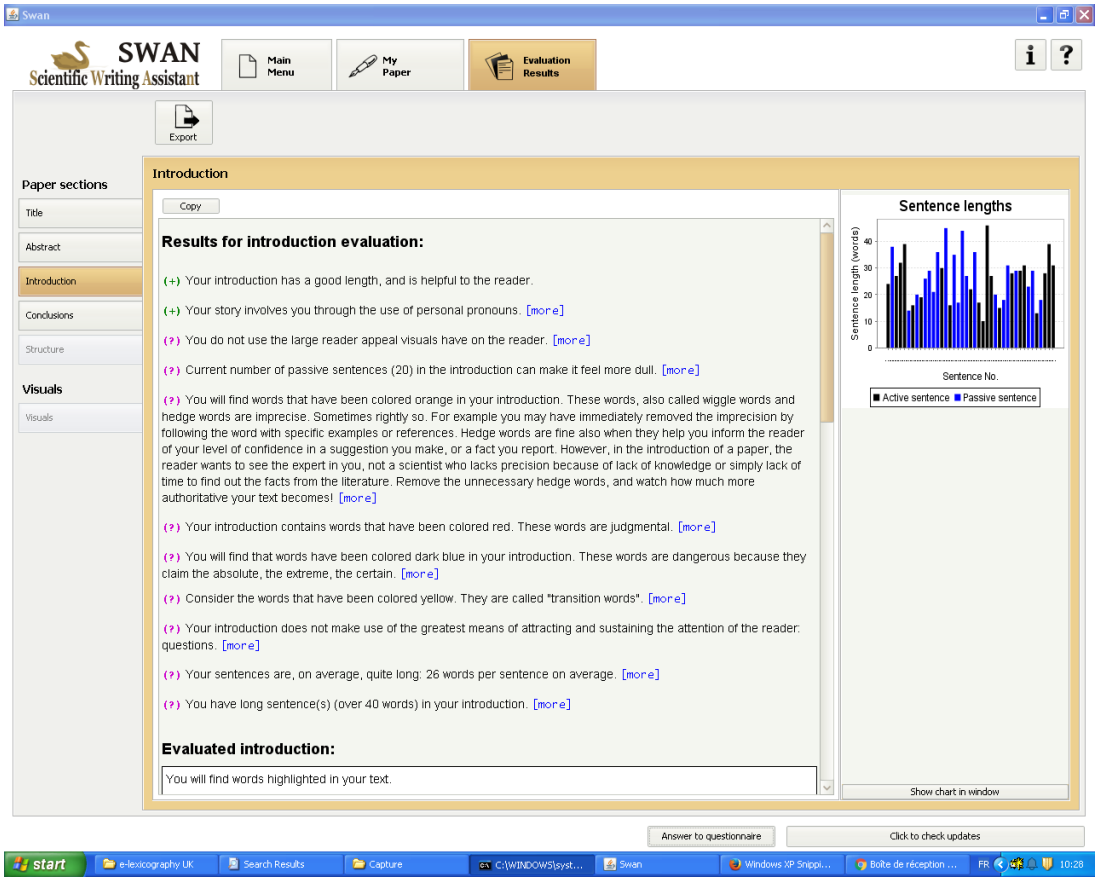


Figure 35 SWAN : Évaluation d'une introduction

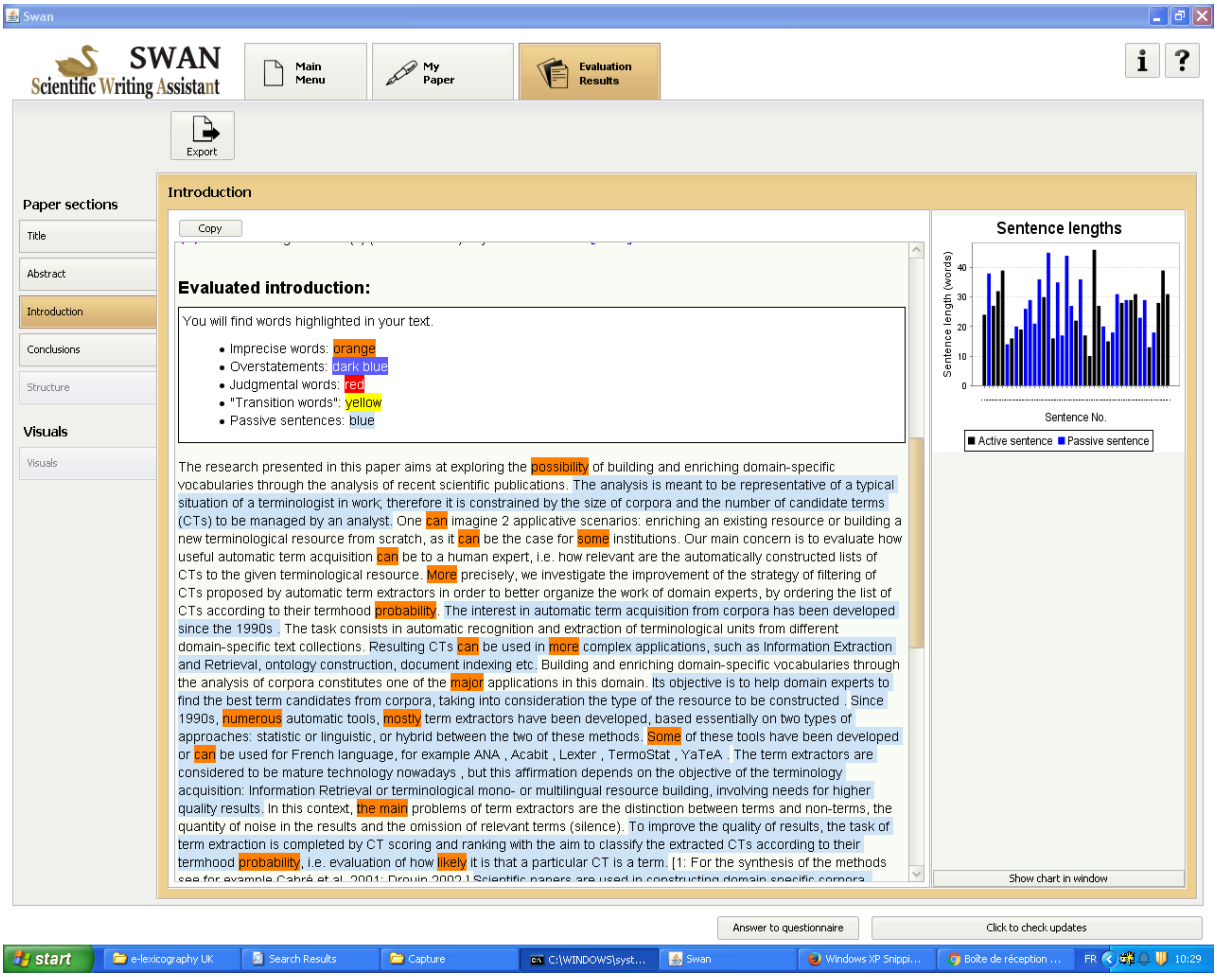


Figure 36 SWAN : Évaluation visuelle d'une introduction

Chaque règle concernant une norme d'écriture se traduit par une mesure pouvant être calculée automatiquement ou semi-automatiquement. Pour donner un exemple précis sur les façons de faire des auteurs du logiciel, la complétude est calculée en nombre de mots contenus dans l'ensemble de l'article par rapport au nombre de mots contenus dans l'introduction (TURUNEN 2013) ; si le ratio est inférieur à 15%, l'introduction peut-être perçue comme incomplète.

En plus des mesures concernant chaque partie d'article, Swan évalue aussi la relation entre les différentes parties, par exemple la consistance entre l'abstract et le titre (sur la présence/ absence de mots-clés dans les deux), ou, sur le même principe, la consistance entre le titre et la structure de l'article.

4.1.2 Enquête sur les habitudes rédactionnelles

Pour nous aider à choisir quelle approche logicielle adopter dans l'accompagnement à la rédaction scientifique pour le biomédical, nous avons conçu une enquête destinée aux professionnels de la Santé[68], avec pour objectifs de :

[68] Diffusée en ligne entre 01/12/2015 et 31/01/2016.

- questionner les habitudes rédactionnelles des chercheurs en sciences biomédicales : comment écrivent-ils les articles scientifiques? Quelles sont les difficultés majeures auxquelles ils se heurtent? Quels outils utilisent-ils pour les surmonter?
- déterminer les types d'outils qui peuvent assister l'écriture, en fonction de leur utilité et de leur faisabilité.

Les 33 questions portaient essentiellement sur les problématiques liées **à la langue et à la façon de rédiger** et non sur **le contenu** des articles. Nous avons aussi veillé à ne pas utiliser de notions linguistiques trop complexes, et à illustrer les questions pour clarifier les concepts utilisées.

Nous avons obtenu 68 réponses :

- 26,5 %[69] des répondants se sont déclarés chercheurs novices (chercheurs rédigeant depuis moins de 3 ans),
- 22% chercheurs confirmés (entre 3 et 10 ans)
- et 51,5% chercheurs chevronnés (plus de 10 ans).

Sur la question du niveau en anglais scientifique, seulement 16% des enquêtés l'a estimé 'pas bon du tout', alors que la grande majorité (51,5%) a opté pour 'assez bon' , 'bon' pour 23,5% et 'très bon' pour 9%.

Sur les questions concernant la langue de rédaction, près de 93% de répondants affirme rédiger les articles directement en anglais, sans passer par le français (70,5 % 'toujours' et 22% 'souvent'). Plus de la moitié fait appel à un traducteur (51,5%), surtout pour la révision du manuscrit en anglais et, beaucoup moins, pour traduire le manuscrit français en anglais (14%).

67,5 % des répondants dit n'avoir aucune formation particulière en rédaction scientifique et avoir appris par soi-même, en rédigeant.

Les questions suivantes portaient sur les difficultés liées à la rédaction, et à l'utilisation d'un sous-ensemble particulier de la langue scientifique. Nous les avons interrogés sur ce qui leur pose problème lors de la rédaction en anglais :

- Utiliser correctement l'anglais scientifique général (Exemples : *This case study confirms the importance of…; The evidence presented thus far supports the idea that …)* ; 48,5% des répondants affirme avoir des difficultés à utiliser correctement l'anglais scientifique général (Figure 37) ;

[69] Pour des raisons de lisibilité, les résultats de l'enquête sont arrondis au demi-point le plus proche, cf. 26,5% pour 26, 41%.

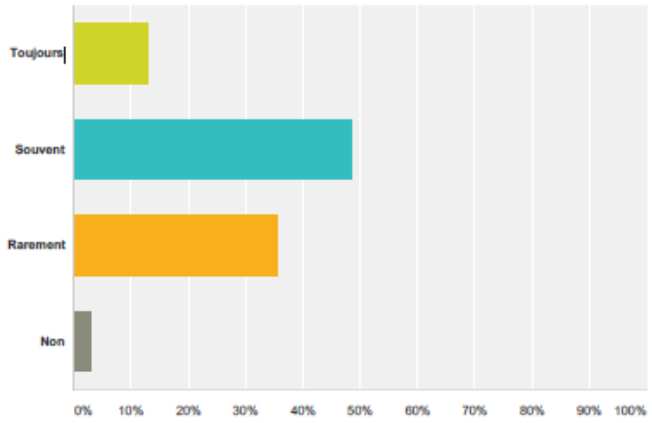


Figure 37 Résultats d'enquête : Difficultés à utiliser l'anglais scientifique général

- Utiliser correctement l'anglais scientifique médical (Exemples : *the parasite acts* (et non : *the parasite behaves*) *; biological markers* (et non : *biological signs*) ; plus de la moitié des répondants affirment avoir des difficultés à l'utiliser correctement (Figure 38 : 41,5% 'souvent' et 13% 'toujours') ;

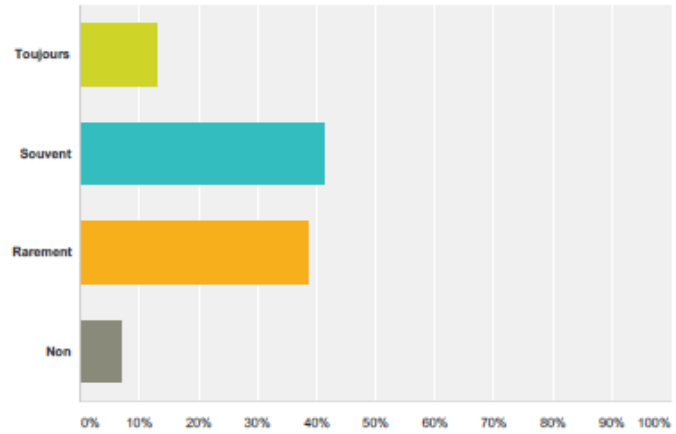


Figure 38 Résultats d'enquête : Difficultés à utiliser l'anglais scientifique médical

- Trouver les équivalents terminologiques des termes français dans leur domaine : 50% des répondants a rarement des difficultés à trouver des termes et 11,5% n'en a jamais (Figure 39) :

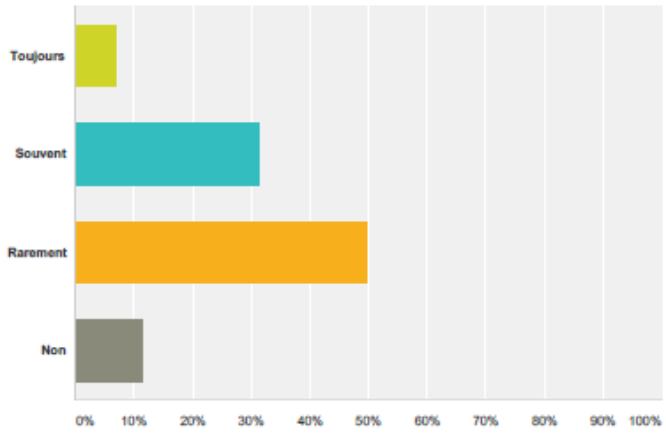


Figure 39 Résultats d'enquête : Difficultés à trouver des équivalents terminologiques

Nous avons détaillé les questions concernant l'utilisation de la terminologie biomédicale pour savoir si les enquêtés avaient des difficultés pour :

- trouver des équivalents de termes, qu'ils soient simples (c'est-à-dire composés d'un seul mot, exemple : *cellule* (fr.) --> *cell* (ang.)) ou complexes (c'est-à-dire composés de plusieurs mots, exemple : *tumeurs épidermoïdes* (fr.) --> *squamous cell tumor* (ang.)) ; 46,5% des enquêtés affirme avoir rarement des difficultés avec les termes, qu'ils soient simples et complexes (Figure 40) ;

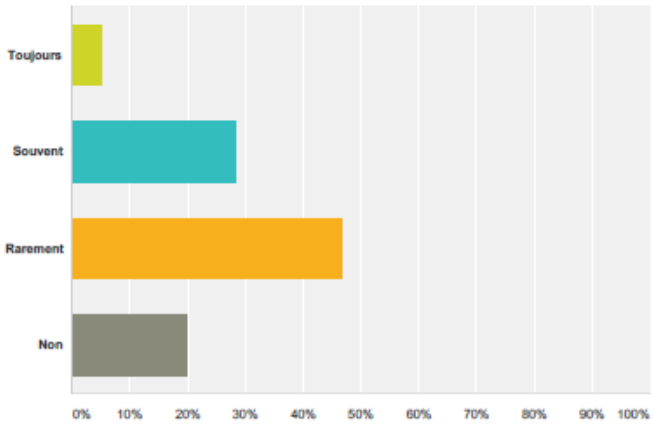


Figure 40 Résultats d'enquête : Termes simples et complexes

- utiliser des termes en contexte, c'est-à-dire, trouver le(s) mot(s) exact(s) qu'on doit utiliser avec un terme dans un contexte donné (exemples : *The patient underwent abdominal surgery...Epidemiology of rabies in animals...Signs and symptoms of the disease at presentation...*) ; 56,5% affirme avoir souvent ce type des difficultés et 13% toujours (Figure 41) ;

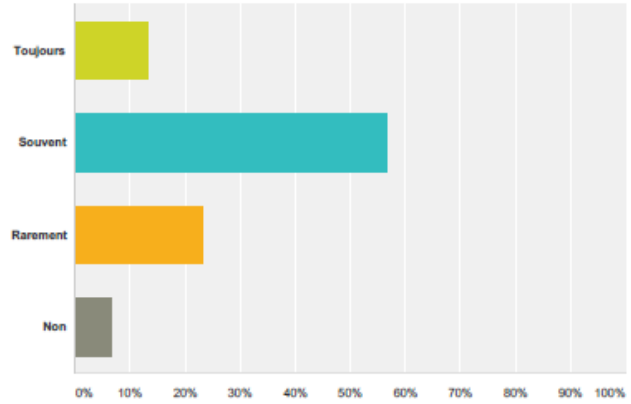


Figure 41 Résultats d'enquête : Termes en contexte

Les questions suivantes portaient sur l'utilisation des outils que l'on peut consulter lors de la rédaction d'articles, sous leurs différentes formes : les dictionnaires, les manuels, les traducteurs automatiques ou les moteurs de recherche.

Nous avons demandé si les enquêtés utilisaient :

- les divers dictionnaires **papier** mono- ou bilingues (Tableau 6) ; le dictionnaire papier le plus utilisé est le dictionnaire bilingue français-anglais de la langue générale (34,85%), mais la plupart des répondants affirme ne pas utiliser aucun dictionnaire papier, qu'il soit monolingue (69,49%) ou bilingue (59,09%) ;

	de la langue générale	de la langue scientifique	terminologique de votre domaine	de synonymes	autres	aucun
Dictionnaires monolingues (anglais-anglais)	25,42%	11,86%	5,08%	13,56%	1,69%	69,49%
Dictionnaires bilingues (français-anglais)	34,85%	15,15%	6,06%	6,06%	3,03%	59,09%

Tableau 6 Résultats d'enquête : Utilisation des dictionnaires papier

- les divers dictionnaires **en ligne** mono- ou bilingues (Tableau 7) ; le dictionnaire en ligne le plus utilisé est, de nouveau, le dictionnaire bilingue français-anglais de la langue générale (83,08%) ; en comparant les deux tableaux (Tableaux 6 et 7), on constate, sans surprise, que l'utilisation de dictionnaires en ligne est nettement supérieure à celle de dictionnaires papier ;

	de la langue générale	de la langue scientifique	terminologique de votre domaine	de synonymes	autres	aucun
Dictionnaires monolingues (anglais-anglais)	56,60%	32,08%	13,21%	32,08%	3,77%	32,08%
Dictionnaires bilingues (français-anglais)	83,08%	43,08%	10,77%	26,15%	3,08%	7,69%

Tableau 7 Résultats d'enquête : Utilisation des dictionnaires en ligne

Nous avons demandé si les enquêtés avaient déjà consulté les manuels pour s'aider lors de la rédaction (Tableau 8). De manière plutôt surprenant, 69,05% affirment n'en avoir jamais consulté.

Choix de réponses	Réponses
Non, je n'ai jamais consulté de manuels d'aide à la rédaction.	69,05%
Oui, j'ai consulté des manuels d'aide à la rédaction en début d'apprentissage, mais je ne le fais plus actuellement.	19,05%
Oui, j'ai consulté et je consulte encore des manuels d'aide à la rédaction.	11,90%

Tableau 8 Résultats d'enquête : Utilisation des manuels

L'utilisation des traducteurs automatiques n'est pas très répandue : 42,65% des répondants affirme les utiliser 'rarement' et 29,41% 'jamais'. Comme exemples de traducteurs

utilisés, ils citent Google Translate, Reverso, Linguee et WordReference[70] (même si ces deux derniers ne sont pas des traducteurs automatiques à proprement parler). Par contre, les moteurs de recherche sont utilisés 'souvent' (57,35%) et 'tout le temps' (33.82%), surtout pour trouver des exemples d'utilisation d'un terme (88,06% ; Tableau 9). En termes d'exemples de moteurs de recherche utilisés, on cite les moteurs de recherche généralistes (Google Scholar, Wikipedia), spécialisés (PubMed[71], Scopus[72], Web of Science[73], ScienceDirect[74]), ainsi que les réseaux sociaux professionnels (LinkedIn, ResearchGate, Academia.edu).

Choix de réponses	Réponses
Pour trouver des références scientifiques.	80,60%
Pour trouver des exemples d'utilisation d'un terme ou d'une expression dans d'autres articles déjà publiés dans votre domaine.	88,06%
Pour trouver des exemples de textes à utiliser comme modèles.	41,79%
Autre.	7,46%

Tableau 9 Résultats d''enquête : Utilisation des moteurs de recherche

Enfin, nous les avons interrogés sur les types d'outils en ligne qui pourraient aider lors de la rédaction, la traduction et la révision. Nous avons défini 5 types d'outils en proposant de les classer de 1 à 5, 1 étant l'outil le plus utile (Tableau 10). Alors que les dictionnaires obtiennent des scores assez médiocres, deux outils semblent être plébiscités : un outil de révision (score de 3,71) et un moteur de recherche de termes et d' « expressions terminologiques » appropriés au domaine (3,67), suivis de près par un moteur de recherche des expressions de la langue scientifique générale (3,19).

	1	2	3	4	5	Score
Un dictionnaire terminologique dans votre domaine, plus complet et /ou plus pertinent que ceux qui existent.	4,76%	9,52%	17,46%	38,10%	30,16%	2,21
Un dictionnaire de la langue scientifique anglaise, plus complet et /ou plus pertinent que ceux qui existent.	3,17%	19,05%	14,29%	23,81%	39,68%	2,22
Un moteur de recherche de termes ou d'expressions terminologiques utilisables dans les articles scientifiques de votre domaine.	28,57%	28,57%	28,57%	9,52%	4,76%	3,67
Un moteur de recherche d'expressions de la langue scientifique générale utilisables dans les articles scientifiques de votre domaine.	15,87%	31,75%	19,05%	22,22%	11,11%	3,19
Un outil de vérification qui pourrait aider à revoir les passages linguistiquement difficiles d'un article.	47,62%	11,11%	20,63%	6,35%	14,29%	3,71

Tableau 10 Résultats d'enquête : Outils à développer

[70] http://www.wordreference.com/fr/
[71] http://www.ncbi.nlm.nih.gov/pubmed
[72] https://www.scopus.com/
[73] http://ipscience.thomsonreuters.com/product/web-of-science/
[74] http://www.sciencedirect.com/

Pour compléter l'enquête, nous avons posé des questions ouvertes sur différents aspects concernant la rédaction scientifique et les outils d'aide. Sur la question concernant les principaux manques ou défauts des outils, on a cité le plus fréquemment :

- manque de spécificité (outils non-adaptés au domaine ; problème de traduction lorsque la terminologie est trop spécifique ; pas assez ciblé anglais scientifique) ;
- manque d'adaptation au contexte de la phrase ;
- absence de système de correction et de vérification, surtout en ce qui concerne la syntaxe de la phrase ;
- perte de temps par des recours à des outils multiples, mais pas vraiment adaptés ;
- défaut de convivialité ;
- méconnaissance des outils en ligne.

Pour résumer, nous avons appris lors de cette enquête que, même si l'époque des dictionnaires papier est révolue, on n'utilise pas encore d'outils en ligne, à part quelques outils généralistes. La préférence va tout de même vers les moteurs de recherche et vers la spécialisation des outils. Nous nous sommes appuyés sur les résultats de cette enquête pour mettre en place les spécifications du logiciel à développer.

4.1.3 Principes de conception

Même si, selon les répondants, rien ne remplacera le rédacteur humain dans le processus de rédaction[75], nous avons pu décider des directions à donner au logiciel d'aide à la rédaction.

Choix de l'approche monolingue

L'approche monolingue (anglais) a été choisie en concertation avec les médecins, s'appuyant sur le fait que la grande majorité des articles scientifiques dans le domaine biomédical sont directement écrits en anglais, sans passer par le français.

Choix de « sous-ensemble » de la langue

Nous avons défini des sous-ensembles de la langue à traiter lors de la rédaction des articles scientifiques : *langue scientifique transdisciplinaire, langue scientifique médicale, terminologie médicale*. **La langue scientifique transdisciplinaire** a été déjà beaucoup traitée et a abouti à différents types de projets et ressources : ARTES, Scientext, Academic Phrasebank développé par John Morley à l'Université de Manchester[76], Academic Word List développé par A. COXHEAD (2000).

En ce qui concerne les deux autres propositions (*langue scientifique médicale, terminologie médicale*), nous nous proposons de mener des travaux en parallèle sur ces deux aspects. Par contre, puisque ces deux notions ne sont pas assez explicites, nous les redéfinissons de la manière suivante.

[75] Ce qui a été commenté de façon très drôle par un des enquêtés : « *Le meilleur outil trouvé à ce jour est Fiona Ecarnot (traductrice médicale IT), c'est le top du top* ».
[76] http://www.phrasebank.manchester.ac.uk/.

Langue scientifique trans-biomédicale

Nous avons renommé la *langue scientifique médicale* en **langue scientifique trans-biomédicale** pour appuyer sur son caractère transversal**.** Nous la définissons comme un ensemble du lexique médical et des collocations autour de ce lexique, présents dans plusieurs sous-domaines du biomédical sous ses formes utilisées/utilisables dans des documents scientifiques. Concrètement, cela veut dire que l'on peut retrouver ce lexique dans n'importe quel sous-domaine du biomédical, puisqu'il n'est pas spécifique à un sous-domaine en particulier.

Exemples

Lexique :
health, medicines, patient, syndrome, disease, to suffer from, side effects, etc.
Collocations :
to deal with the syndrome, suffer from a syndrome, to relieve symptoms of, etc.

Les premiers résultats de ces travaux sont décrits dans le & 4.1.4.1.

Terminologies des sous-domaines biomédicaux

Plutôt que de parler de la *terminologie médicale*, nous proposons le terme de *terminologies des sous-domaines biomédicaux*, cette fois-ci pour appuyer sur le caractère spécifique au sous-domaine de chaque terminologie.

La terminologie d'un sous-domaine biomédical diffère du lexique trans-biomédical en ce qu'elle est caractéristique d'un sous-domaine précis (et non commune à l'ensemble des sous-domaines du biomédical).

Nous avons distingué 3 types de problématiques concernant la terminologie :

- Recherche de termes simples (composés d'un mot) :

hyperglycemia, cardiomyopathy, immunosuppression

- Recherche de termes complexes (composés de plusieurs mots) :

bicuspid aortic valve, alveolar echinococcosis, chronic inflammatory diseases

- Contextualisation de termes simples et complexes (Termes en contexte) :

to contract echinococcosis, alveolar echinococcosis in animals, diagnosis and treatment of alveolar echinococcosis

Selon l'enquête, la problématique la plus répandue concerne la contextualisation de termes. Lorsqu'il s'agit de trouver le mot exact que l'on doit utiliser avec un terme, 57% des répondants déclarent avoir 'souvent' un problème à trouver le mot exact qui va avec un terme et 13% 'toujours'. Parallèlement, les répondants préfèrent aussi un outil de type 'moteur de recherche' que de type 'dictionnaire' (cf. & 4.1.2).

Par conséquent, nous projetons de construire un moteur de recherche de termes en contexte, composé de deux parties :

- Une partie 'dictionnaire', qui permet une recherche par terme (entrée du dictionnaire), dans lequel nous aurons préétabli et validé des relations collocationnelles pour les termes ;
- Une partie 'moteur de recherche', qui permet une recherche par terme mais aussi par tous les autres mots du texte, et qui permettra d'afficher les exemples de phrases avec les termes en contexte.

Lorsque l'utilisateur formulera une recherche, elle sera envoyée aussi bien au dictionnaire qu'au corpus ; c'est-à-dire que l'on pourra afficher à la fois les réponses du dictionnaire (s'il y en a) et une liste des phrases du corpus. Ces fonctionnalités existent déjà dans les logiciels de type Linguee ou TradooIT, mais les différences principales avec ces logiciels de référence sont les suivantes :

- Pas de **traduction**, mais :
    - **La viabilité de sources** : contrairement à Linguee et TradooIT, les corpus de base ne seront composés que d'articles publiés dans des journaux reconnus (pour garantir la qualité de l'information) ;
    - **La spécialisation** : lors de nos différentes consultations avec les médecins, nous avons compris qu'il serait utile de construire des outils très spécialisés, non seulement par rapport au grand domaine du biomédical, mais aussi à un sous-domaine précis. Le domaine test que nous avons décidé d'explorer est *Essais cliniques* qui puisera dans le corpus de 3 sous-domaines : *Essais cliniques en cardiologie, Essais cliniques en gastroentérologie et Essais cliniques en pneumologie*. Par conséquent, le lexique terminologique contextualisé sera un lexique commun à ces 3 sous-domaines. Par la suite, pourront être développés des lexiques propres à chacun de ces sous-domaines, et d'autres sous-domaines de spécialités médicales (cancérologie, rhumatologie…) ou ciblés sur le type de thérapeutique concernée (médicaments, dispositifs médicaux, interventions invasives ou non-invasives…).
- **La présentation des termes avec des « fiches de contextualisation »** : les contextes d'utilisation et collocations possibles pour chaque terme seront présentés de manière structurée (en plusieurs groupes, chaque groupe contenant des exemples similaires) et non pas sous forme d'une liste non-ordonnée.

Pour construire un tel outil, il y a plusieurs problématiques à traiter :

- **Corpus** : Nous allons utiliser 2 corpus : le corpus PLOS pour le lexique trans-biomédical (cf. &4.1.4.1) et un autre corpus que nous avons construit à partir des meilleurs journaux des spécialités concernées (au niveau de la qualité scientifique et éditoriale) pour le domaine d'application *Essais cliniques* (cf. &4.1.4.2).
- **Recherche en corpus (du point de vue de l'utilisateur)** : comment les utilisateurs, non-linguistes vont rechercher l'information à partir de cet outil (vu que ce sera un outil monolingue) ? Quelles stratégies va-t-on mettre en place si on ne sait pas comment

dire en anglais '*présentation*' dans le contexte d' '*antigène*' (***'Trois types de cellules ont constitutionnellement des propriétés de présentation de l'antigène …') ou 'déclencher'*** dans le contexte de *'réponse immunitaire'* ?

- **Recherche en corpus (du point de vue du concepteur)** : quels types de recherche sont les plus pertinentes pour aboutir à un résultat ? Doit-on lemmatiser, POS-tagger le corpus ? Permettre des recherches sur les lemmes ? des recherches approximatives ?
- **Établissement des terminologies en contexte** : Quel que soit le sous-domaine, il faut d'abord établir la terminologie du sous-domaine, pour pouvoir ensuite contextualiser les termes (simples et complexes). De nouveau, comme pour le lexique trans-biomédical, se pose la problématique de :
    - **Méthodologie de l'établissement de la terminologie** (à partir du corpus ? en s'appuyant sur un dictionnaire du domaine ?) ; combien de termes retient-on pour le dictionnaire ? Quels sont les outils que l'on utilise ? Est-ce que l'on identifie les termes complexes et comment ?
    - **Contextualisation** : comment fait-on (critères, outils) ? que retient-on ? Comment organise-t-on la 'fiche de contextualisation' ? Doit-on POS-tagger ?
    - **Exemples** : comment trouve-t-on les meilleurs exemples dans le corpus ?
- **Visualisation** : comment présente-t-on les résultats ? Comment présente-t-on un réseau collocationnel ? Ceci nous parait très important, vu que l'on risque d'avoir beaucoup d'informations sur un terme en contexte.
- **Validation expert** : À quel moment ? Sur quoi ?
- **Evaluation** : Comment évalue-t-on le logiciel du point de vue du résultat et de son utilité comme aide à la rédaction ?

4.1.4 Premiers résultats de nos recherches

Deux directions de recherche sont développées parallèlement pour aboutir au système d'aide à la rédaction scientifique.

Premièrement, nous travaillons sur la notion de la langue scientifique trans-biomédicale[77] : c'est une notion qui n'existe pas vraiment dans la littérature, nous posons donc les bases de sa définition. Ce travail s'accompagne d'une réflexion sur la représentions des termes contextualisés.

Deuxièmement, nous travaillons sur le recensement et la description du lexique du domaine en contexte. Pour ce faire, nous construisons un corpus de référence[78] pour bâtir la

[77] Ce travail est confié à Anastasia GALMICHE, dans le cadre de son mémoire de Master TAL, sous ma direction.
[78] Cette tâche a été effectuée par François-Claude REY lors d'un stage de Master 2 sous une direction conjointe avec I. ATANASSOVA.

terminologie des essais cliniques et disposer d'exemples en corpus. Nous présentons les premiers résultats de ces recherches dans les deux paragraphes qui suivent.

4.1.4.1 Lexique trans-biomédical contextualisé

L'objectif consiste à concevoir un dictionnaire contextualisé du lexique trans-biomédical utilisé dans les articles scientifiques et accompagné d'un moteur de recherche en contexte d'exemples.

Le lexique trans-biomédical sera constitué des lexies spécialisées communes à l'ensemble des sous-domaines biomédicaux. Il existe deux avantages à constituer un tel lexique :

- d'une part, l'établissement d'un tel lexique serait fait une fois pour toutes ; de par sa transversalité, il serait intéressant pour tous les sous-domaines du biomédical et pourrait être intégré à n'importe quel outil très spécialisé, consacré par exemple à la cardiologie ;
- d'autre part, lors de l'établissement des terminologies des sous-domaines, il ne sera plus nécessaire de travailler sur le lexique trans-biomédical.

Pour définir le lexique trans-biomédical nous nous sommes appuyés sur le concept de lexique scientifique transdisciplinaire qui est constitué des lexies servant à décrire les activités scientifiques et la méthodologie de la recherche à travers les différentes disciplines.

Les premières listes de vocabulaire transdisciplinaires ont été créées dans les années 1970 : il s'agissait de l'Academic Vocabulary List (CAMPION et ELLEY 1971) et l'American University Word List (PRANINSKAS 1972). Les chercheurs se sont appuyés sur les critères de fréquence et de répartition des mots dans un corpus multidisciplinaire pour inclure des lexies dans ces listes. GHADESSY et LYNN (GHADESSY 1979; LYNN 1973) ont choisi de regrouper dans leurs listes les mots les plus fréquemment annotés par les étudiants dans les manuels, une annotation représentant la difficulté à retenir un mot considéré comme important pour un étudiant. Ces quatre listes ont ensuite été regroupées dans la University Word List (XUE et NATION 1984) qui au total comprend 800 mots représentatifs[79].

La liste la plus connue est l'Academic Word List, établie dans les années 2000 par COXHEAD (2000) à l'aide d'un corpus de 3 500 000 mots provenant de manuels universitaires portant sur 28 disciplines différentes. Cette liste est constituée des 570 mots représentatifs de familles de mots ayant une fréquence minimum de 100 et qui sont utilisés dans au moins 14 domaines différents. Ces 570 familles de mots couvrent 10% des mots du corpus (Tableau 11).

Analysis	Defintion	Indicate	Procedure
Approach	Derived	Individual	Process
Area	Distribution	Interpretation	Required
Assessment	Economic	Involved	Resard
Assume	Environment	Issues	Response
Authority	Established	Labour	Role

[79] Dans ces listes un mot représente toujours une famille dérivationnelle de mots ; par contre ce n'est pas toujours un lemme, cela peut être la forme la plus fréquente dans le corpus.

Available	Estimate	Legal	Section
Benefit	Evidence	Legislation	Sector
Concept	Export	Major	Significant
Conistent	Factor	Method	Similar
Constitutional	Financial	Occur	Cource
Context	Formula	Percent	Specific
Contract	Function	Period	Structure
Create	Identified	Policy	Theory
Data	Income	Principle	Variable

Tableau 11 Premiers mots dans l'AWL (COXHEAD 2000)

Au fil des années, ont émergé des listes académiques basées sur une discipline unique, telles que l'informatique (LAM 2001), les affaires (HSU et al. 2011), l'ingénierie (MUDRAYA 2006), l'agriculture (MARTINEZ, BECK et PANZA 2009), le journalisme (CHUNG 2009) ou bien la théologie (LESSARD-CLOUSTON 2006).

Pour la médecine, la Medical Academic Word List (MAWL) a été proposée par WANG *et al*. (2008). Cette liste réunit 623 mots communs aux sous-disciplines biomédicales les plus représentées dans les écrits biomédicaux (Tableau 12). Elle est construite à partir d'un corpus composé de 288 textes provenant de 32 disciplines médicales et écrits par au moins un auteur anglophone. Le corpus comprend au total 1 093 011 mots appartenant à 31 275 familles de mots différentes.

Numéro	Mot représentatif	Numéro	Mot représentatif	Numéro	Mot représentatif
1	cell	11	tissue	21	therapy
2	data	12	dose	22	indicate
3	muscular	13	gene	23	area
4	significant	14	previous	24	obtain
5	clinic	15	demonstrate	25	research
6	analyze	16	normal	26	vary
7	respond	17	process	27	activate
8	factor	18	similar	28	require
9	method	19	concentrate	29	induce
10	protein	20	function	30	cancer

Tableau 12 Premiers mots de la MAWL (WANG et al. 2008)

La méthodologie s'appuie sur les critères utilisés par COXHEAD pour la création de l'Academic Word List, en y ajoutant le critère de spécificité. Ainsi, pour qu'une famille de mots soit retenue dans la MAWL, elle doit répondre aux critères suivants :

- elle doit être présente dans au moins 16 sous-domaines médicaux ;
- elle doit apparaître plus de 30 fois dans le corpus ;
- elle doit être spécifique au domaine médical.

Afin de vérifier le dernier critère, WANG a fait appel à des experts du domaine biomédical. De manière assez surprenante, ils ont décidé de supprimer 27 familles de mots car jugés trop spécialisées (Tableau 13).

Numéro	Mot représentatif	Numéro	Mot représentatif	Numéro	Mot représentatif
1	pathogenesis	11	posterior	21	Ischemia
2	cytokine	12	anterior	22	Cerebral
3	epithelial	13	lysis	23	Dorsal
4	mitochondrial	14	cardia	24	Hemmorhage
5	carcinoma	15	necroisis	25	Pathophysiology
6	ligand	16	cutaneous	26	exogenous
7	situ	17	stent	27	Phenotypic
8	lymphoid	18	vivo		
9	vitro	19	hepatic		
10	pulmonary	20	aortic		

Tableau 13 MAWL : 27 familles de mots supprimées par les experts

De l'autre coté la liste contient des mots tels que *whereas* and *thereby*, qui ne sont, de toute évidence, pas liés au domaine. Le problème de la liste de WANG réside dans le fait que l'appartenance d'un mot à la liste est décidée sur le mot lui-même sans avoir recours au contexte. Or, comme le constate FRASER (2007, 2009a, 2009b), la majorité des mots retenus selon les critères de fréquence et de répartition (donc, les critères de WANG) sont ambigus du point de vue de leur spécialisation dans le domaine. Il existe des mots qui ont pour origine un sens spécialisé mais qui sont communément utilisés dans le lexique général ; d'un autre côté, les termes crypto-techniques sont des mots du lexique général ou transdisciplinaire qui acquièrent un sens spécialisé dans un domaine. Fraser décide donc de procéder à une étude des cooccurrents afin de déterminer le sens véhiculé par l'unité lexicale analysée.

L'étude des cooccurrents est nécessaire pour la détermination du sens mobilisé par une unité lexicale. En effet, si une unité lexicale a pour cooccurrents des termes du domaine étudié, alors il est fort probable qu'elle représente elle-même un terme de ce domaine. Ainsi, Fraser a déterminé que le verbe *block* était utilisé dans le sens « empêcher l'action d'une drogue » car les collocations fréquentes dans lequel *block* apparait *sont blockade of [receptor], channel blocker(s), beta blocker(s).*

Comme la liste de Wang est la seule liste du lexique médical académique qui existe à ce jour, nous nous en sommes servis tout d'abord pour vérifier la pertinence du lexique retenu sur un corpus plus grand, à savoir le corpus PLOS[80]. C'est un corpus de 46 000 articles extraits de 5 revues PLOS différentes : PLOS Biology, PLOS Computational Biology, PLOS Medicine, PLOS Pathogens et PLOS Neglected Tropical Diseases. C'est aussi ce corpus qui nous sert de base pour nos recherches sur le lexique trans-biomédical : nous allons l'utiliser pour contextualiser le lexique retenu afin de construire le dictionnaire, ainsi qu'il sera une base des exemples pour le moteur de recherche sur la langue trans-biomédicale.

L'expérimentation de Wang a été adaptée au corpus PLOS[81], qui contient 41 millions de mots et que nous avons divisé en 46 sous-domaines. Par conséquent, les critères ont été réajustés : pour la répartition, un mot doit appartenir à au moins 23 domaines et pour la fréquence, avoir une fréquence d'au moins 30 occurrences pour 1 millions de mots (donc 1230 occurrences dans l'ensemble du corpus).

Sur 623 familles de mots de la liste de WANG, 53 familles (8%) n'ont pas respecté ces critères. Le critère non tenu est plutôt celui de la fréquence (53 familles) que de la répartition (3 familles : *ration, perception* et *append*). Parmi les 53 familles qui ne respectent pas le critère de la fréquence, la majorité n'est pas spécifique au biomédical : *concomitant, aknowledge, comment*, *inferior, thereafter etc.*, mais certaines correspondent bien à des lexies spécialisés : *methanol, catheter, laser* etc. En plus, la répartition de mots selon la fréquence ne correspond pas dans les deux expérimentations : on ne retrouve pas les termes les plus et les moins fréquent aux mêmes rangs dans les deux listes. Par exemple, si on fait la comparaison entre les 50 familles de mots les moins fréquentes dans les deux corpus, on retrouve seulement 19 familles en commun. Dans les résultats les plus fréquents, on retrouve 27 familles en commun dans les premiers 50 familles de chaque liste. En général, plus on va vers les résultats fréquents, plus les mots correspondent à des lexies spécialisés (*cell, gene, infect*, etc.).

Enfin pour conclure, cette évaluation démontre que si nous voulons prendre appui sur la MAWL pour le lexique trans-biomédical, il faut non seulement faire un tri entre lexies spécialisées et non-spécialisées mais aussi supprimer les termes qui n'ont pas une fréquence suffisante dans un corpus plus large. Après avoir enlevé les termes insuffisamment fréquents et éliminé les termes non spécifiques au domaine biomédical à l'aide de dictionnaires

[80] Nous remercions l'OST, Montréal pour nous avoir donné le droit d'utiliser ce corpus.
[81] Ce travail, effectué par Anastasia GALMICHE dans le cadre de son mémoire de Master TAL (sous ma direction), n'a pas été encore publié.

spécialisés, nous nous retrouvons donc avec 320 familles de mots retenues pour le lexique trans-biomédical[82].

Il en reste que la liste n'est composée que des termes simples alors que la terminologie médicale est riche en termes composés, que nous voulons aussi intégrer au lexique trans-biomédical.

La suite de travaux consiste en une contextualisation de lexies spécialisées retenues dans le corpus, avec trois problématiques majeures :

- dictionnaire : définir la « fiche de contextualisation » du point de vue de la forme et du fond : quelles informations retient-on et comment les présente-t-on à l'utilisateur ?
- cooccurrents : définir la méthodologie de contextualisation des lexies spécialisées pour obtenir leur cooccurrents ;
- moteur de recherche : réfléchir au moyen de trouver dans le corpus les exemples les plus pertinents.

4.1.4.2 Terminologie contextualisée des essais cliniques

Les travaux concernant la construction de la terminologie en contexte des essais cliniques ont débuté par la construction d'un corpus spécialisé adapté. Pour construire ce corpus, nous avons pris en compte plusieurs critères :

- La représentativité :

Afin d'évaluer et de valider la représentativité de journaux à prendre en compte pour constituer le corpus, nous avons fait appel à des experts (médecins et traducteurs en biomédical). Pour chaque sous-domaine du projet (cardiologie, gastro-entérologie et pneumologie), nous avons choisi quelques revues de haut niveau scientifique et langagier : il était très important de disposer d'articles avec un très bon niveau de rédaction en anglais. Pour compléter, nous avons choisi plusieurs journaux généralistes du domaine des essais cliniques. Les détails de chaque journal ont été consignés selon les indications suivantes (Tableau 14) :

Identifiant	Nom donné au journal, en 4 lettres minuscules. La 1ère lettre correspond à la spécialité du journal : - *Clinical Essays* → ggas, ggan, gjoh, ghep, prcc, pcho, pchs, ptho, cacc, ccir, ecid, elan et ejam.
Nom du journal	Nom en clair du journal.
Année de création	Année de création du journal.
Sous-domaine	Sous-domaine biomédical.
Périodicité	Périodicité de parution du journal pour les derniers exemplaires.
Nombre moy. aprox. d'articles par numéro	Approximation du nombre d'articles scientifiques originaux dans chaque édition du journal, pour les derniers numéros.

[82] Voir les 100 familles les plus fréquentes dans l'Annexe 9.4.

Nombre moy. aprox. d'articles par année	Approximation du nombre d'articles scientifiques originaux dans la dernière année de parution du journal.
Formats	Type de HTML téléchargé, c'est à dire la structure interne qui lui a été donnée par l'éditeur (Eselvier, ...).
Années dans le corpus	Indication abrégée de la période exacte des articles téléchargés, pour retrouver quels sont les articles contenus dans le corpus.
Années dans le corpus (détails)	Indication détaillée de la période exacte des articles téléchargés.
Nombre d'articles dans le corpus	Nombre d'articles scientifiques originaux téléchargés dans ce corpus.
Formats téléchargés	Formats des fichiers des articles téléchargés. Ces fichiers se trouvent dans le dossier ayant l'identifiant du journal (ex. : 'ggas.zip'), dans les sous-dossiers 'table', 'html' et 'pdf' (et 'txt' pour quelques journaux).
Formats générés	Cet emplacement est vide pour l'instant. Il indiquera les formats des fichiers générés par le traitement des articles à partir de leurs formats HTMLs originaux téléchargés "bruts".
Rubrique des articles	Indication présente dans le journal et sur laquelle est basée le choix des articles téléchargés (ex. : 'original article').
Site web	URL de la page web d'accès aux articles téléchargés.
Droits	URL de la page web ou du PDF d'information sur les droits d'usage des articles téléchargés.

Tableau 14 SARS : Description des journaux dans le corpus

Nous avons donc sélectionné au total 11 journaux de spécialité pour constituer le corpus de référence : 3 revues en gastroentérologie, 3 revues en pneumologie, 2 revues en cardiologie et 3 revues généralistes en essaie cliniques. Il était aussi important de disposer des articles récents pour qu'ils soient représentatifs d'une terminologie à jour (donc les années 2014 à 2016). Enfin, nous n'avons retenu que des articles originaux, c'est-à-dire des articles provenant de la rubrique « Recherche »des journaux concernés.

- La taille :

Pour s'assurer d'une taille de corpus suffisante, nous avons téléchargé au moins 100 articles par journal. Bien que nous ne disposons pas pour le moment des statistiques précises sur ce corpus, au moins 3 500 fichiers ont été manipulés pour le constituer (dans des différents formats, voir plus bas).

- L'accessibilité et le(s) format(s) :

Il s'agissait de considérations plus techniques, à savoir la possibilité d'avoir un accès soit à l'intégrité des articles scientifiques, soit à des articles en nombre suffisant (au moins 100), dans des formats exploitables par la suite. Chaque article a été sauvegardé en HTML et PDF (exceptionnellement .txt). Nous avons aussi pris en compte les droits d'utilisation, préférant télécharger les articles que nous pourrions utiliser librement pour la recherche et l'enseignement.

Dans l'état d'avancement des travaux, le corpus finalisé est brut, c'est à dire que les formats de documents originaux sauvegardés sont différents en fonction du journal (ex. : les structures des articles ne sont pas les mêmes pour tous les journaux, qu'il s'agisse des textes eux-mêmes ou du balisage HTML). Pour rendre le corpus plus utilisable, une étape ultérieure de transformation est nécessaire : les articles devront être convertis dans un standard pivot, à définir à l'aide de la norme XML DocBook[83]. Ce passage à un même standard pratique facilitera la mise en œuvre de traitements automatiques du corpus.

4.2 Plateforme d'apprentissage du lexique spécialisé

Concevoir des environnements informatiques pour *enseigner* (et non pour apprendre) semble un objectif nouveau, même si le besoin de donner à l'enseignant le rôle d'utilisateur principal des logiciels pédagogiques a déjà été identifié (RIOT 2004, PHO 2015)[84]. Cependant, il existe très peu d'outils logiciels centrés sur les besoins des enseignants de langues étrangères et encore moins de langues étrangères de spécialité. Pourtant, le besoin existe comme en témoigne le développement rapide de disciplines telles que LANSAD (*Langue pour Spécialistes d'Autres Disciplines*) ou VOLL (*Vocationally (and Professionally) Oriented Language Learning*).

Il existe des systèmes de création d'exercices pour l'apprentissage des langues étrangères : MALAFEEV (2015) en énumère 15, portant sur les différentes langues (principalement anglais, mais aussi russe, basque et arabe), et comportant différents types d'exercices (phases à trous, recomposition de phrases, questions à choix multiples, etc.). Pourtant, aucun d'eux n'est adapté à l'apprentissage des langues spécialisées, ce qui prouve qu'elles n'ont pas encore attiré l'attention de chercheurs en ALAO (Apprentissage des Langues Assisté par Ordinateur). Nous n'avons d'ailleurs trouvé aucun système automatique d'aide à la création d'exercices d'apprentissage dédiés aux langues spécialisées. Partant de ce constat, nous nous sommes fixé comme objectif de concevoir une plateforme d'aide semi-automatisée à la création des matériels didactiques de langues étrangères de spécialité. Comparativement aux autres logiciels qui peuvent servir à produire des matériels didactiques pour les langues, nous nous orientons vers les objectifs suivants :

- mettre au point une plateforme qui répond d'abord aux besoins des enseignants (plutôt qu'à ceux des apprenants qui n'en sont pas utilisateurs directs mais bénéficiaires) ;
- prendre en compte des particularités des langues spécialisées de niveau académique dans la production didactique assistée par ordinateur ;
- structurer l'aspect générique de la plateforme, pour qu'elle puisse être réutilisable pour d'autres langues et d'autres spécialités.

Nous décrivons dans la suite de cette section la première étape dans la construction de cette plateforme, à savoir la conception et la mise en place d'un générateur automatique d'exercices d'apprentissage du vocabulaire spécialisé.

[83] www.docbook.org
[84] Cette section est basée sur Rey FC et al. (2016a et 2016b).

Le fonctionnement que nous avons imaginé pour cet outil est le suivant. La plateforme dispose d'un vocabulaire spécialisé et d'un corpus de textes dans le domaine étudié. L'enseignant apporte un nouveau texte (que nous appellerons 'texte de référence') dans le même domaine à partir duquel il souhaite créer des exercices de vocabulaire. Il choisit les termes à étudier, soit à partir du texte, soit à partir de la liste du vocabulaire. Ces termes sont ensuite recherchés dans le texte de référence et dans le corpus, pour fournir à l'apprenant d'autres exemples d'utilisation du terme à apprendre que ceux du texte de référence. Ceci a pour objectif d'illustrer l'utilisation du même terme dans d'autres contextes, ainsi que d'aider l'apprenant à trouver le bon terme et renforcer l'apprentissage. Ainsi l'apprentissage est basé sur les exemples tirés d'un corpus de textes authentiques.

Pour mettre en place notre expérimentation, nous avons choisi le thème d'un cours de Master donné à l'Université de Franche-Comté, *English for Geographers*.

4.2.1 État de l'art de la génération automatique d'exercices

A la jonction de l'informatique et de l'apprentissage des langues, le secteur de la génération automatique d'exercices apparaît dispersé quant aux disciplines d'origine des auteurs de publications (professorat, pédagogues, sociétés privées, informaticiens, TAL, etc.). Par ailleurs, les appellations hétéronymes concernant les disciplines et les concepts quasi-communs du secteur sont nombreuses, allant de « logiciel d'édition de contenu pédagogique » à « environnements informatiques pour l'apprentissage humain ». Cette pluralité semble découler du fait que la génération automatique d'exercices est une problématique encore jeune et en pleine expansion.

Il n'existe pas de plateforme dédiée à l'apprentissage de langues spécialisées. Au niveau de l'enseignement universitaire, on trouve deux plateformes dédiées à la génération d'exercices, MIRTO et ASKER. Le projet MIRTO (ANTONIADIS et al. 2005) aborde les problématiques didactiques des enseignants de langues, en se centrant sur la création semi-automatique d'exercices de langue générale qui peuvent se succéder sur la plateforme, de manière prédéfinie par un enseignant, pour composer des scénarios qui tiennent compte des réponses des apprenants. La plateforme ASKER (LEFEVRE et al. 2015), elle, est très généraliste : elle sert de support à la création d'exercices dans n'importe quelle matière enseignée à l'université[85]; et, de ce fait, elle n'intègre pas de connaissances relatives aux domaines, lesquelles doivent être apportées par un enseignant.

En ce qui concerne les communautés scientifiques autour de l'enseignement des langues spécialisées, aucune ne fait mention de logiciels dédiés à l'apprentissage des langues spécialisées du niveau universitaire ou de logiciels dédiés à l'apprentissage des langues étrangères de spécialité. Que l'on regarde les publications du Groupe d'Étude et de Recherche en Anglais de Spécialité (GERAS) rassemblées dans la revue Anglais de Spécialité (ASP) ou les publications du Groupe d'Étude et de Recherche en Espagnol de Spécialité (GERES), c'est le même constat. . Les logiciels qui se rapprochent le plus de ces objectifs sont ceux qui aident à

[85] Par exemple, elle est actuellement utilisée pour l'enseignement de l'informatique.

l'apprentissage des langues générales au niveau élémentaire dans les filières d'enseignement à l'université (TANO 2011).

Génération d'exercices de vocabulaire et vocabulaire de spécialité

Les exercices d'apprentissage des langues peuvent être divisés en 2 catégories (PEREZ-BELTRACHINI et al. 2012) : les exercices basés sur des phrases réelles (« real life sentences », c'est-à-dire les phrases extraites de documents existants), et les exercices basés sur une syntaxe et un vocabulaire limités. Notre travail se situe dans la première catégorie, celle de phrases tirées des documents authentiques, puisque l'apprentissage du terme est aussi important que l'apprentissage de l'environnement dans lequel il est naturellement employé. Il existe plusieurs exerciseurs de ce type (MALAFEEV 2015), mais aucun ne concerne les langues spécialisées.

Thierry SELVA (2002) décrit plusieurs types d'exercices d'apprentissage de la langue générale dans le cadre du projet ALFALEX (Environnement d'apprentissage lexical interactif pour apprenants du français). Il décrit notamment les exercices sur les collocations, où un ensemble de phrases est sélectionné dans un corpus pour illustrer les collocations les plus fréquentes. Pour construire les exercices, une partie de la collocation est affichée et l'autre cachée. Le but consiste à compléter la collocation à partir de sa partie affichée et du reste du contexte de la phrase. Le système accepte plusieurs réponses contenant des nuances sémantiques (verbes alternatifs, intensification, etc.) en s'appuyant sur le dictionnaire en ligne DAFLES (Dictionnaire d'Apprentissage du Français Langue Étrangère ou Seconde).

Pour les langues spécialisées et le langage technique, CHARNOCK (1999) propose de travailler sur des textes courts (résumés, introductions d'articles, etc.) pour tenir compte des apprenants qui n'ont pas de bases en langue étrangère encore bien établies, tout en supposant qu'ils ont des connaissances adéquates dans la discipline. Il suggère aussi l'impossibilité d'un travail efficace sur la langue sans la prise en compte du contexte et des intentions communicatives. Il signale aussi que les textes authentiques de certaines disciplines comportent souvent des archaïsmes linguistiques qui compliquent la tâche des apprenants.

Selon MALAFEEV (2015), un système de génération d'exercices basé sur des listes de mots doit prendre en compte des facteurs tels que les majuscules, l'orthographe, la ponctuation, la longueur des mots, la distance entre mots à trouver, le nombre des mots dans le texte, la longueur des mots, etc. Certaines règles établies à partir de ces facteurs permettent de lever des ambiguïtés que l'usage des seuls dictionnaires ne résout pas. D'après GUREVYCH et al. (2009 : 11), pour l'apprentissage des langues, les paramètres de sélection des mots à remplacer par des blancs dans les exercices à trous peuvent être :

- chaque n^{ième} mot dans le texte ;
- la fréquence des mots ;
- des mots appartenant à des parties du discours, tels que les noms, verbes, adjectifs et adverbes, et dont le sens peut être ciblé ;
- des mots obtenus par un apprentissage automatique basé sur un ensemble de questions saisies (*input questions*) utilisées comme données d'apprentissage.

Pour la création de tests de vocabulaire par les enseignants, COOMBE (2011) considère le problème du format : le test est valide si les apprenants ont l'expérience du format de présentation du contenu, s'il n'y a pas d'ambigüité sur comment répondre et comment interpréter les réponses, et si le format a un effet positif sur l'apprentissage, par exemple en aidant la répétition ou l'extension du vocabulaire.

Les « contextes riches »

Dans un exercice à trous, les termes sur lesquels porte l'exercice sont remplacés par des blancs dans les phrases proposées aux apprenants. Il est donc nécessaire que chaque phrase permette de deviner le terme manquant. Pour la préparation des exercices, pour chaque terme recherché, il faut être en mesure d'identifier automatiquement dans les textes apportés par les enseignants des phrases riches, c'est-à-dire des phrases avec un contexte assez riche en informations pour que l'apprenant puisse restituer les termes manquants. Le concept de « contexte riche » est donc important, puisqu'il aide à définir, localiser et prendre en compte les informations contextuelles pertinentes lors de la génération de l'exercice, et ensuite à faciliter la résolution de l'exercice pour un apprenant.

Firas HMIDA *et al*. (2015) proposent de mettre en œuvre la notion de Contextes Riches en Connaissances (CRC) introduite en 2001 par Ingrid MEYER (2001) « pour désigner les contextes qui illustrent des relations entre les termes d'un domaine spécialisé ». Ils proposent l'extraction de 'contextes conceptuels et linguistiques' dans les corpus monolingues spécialisés et dans un corpus scientifique de volcanologie selon deux méthodes :

- la première méthode s'appuie sur la présence du terme à illustrer et l'exploitation d'indices lexicaux pour extraire, grâce à des marqueurs de relations conceptuelles entre termes, des contextes riches en connaissances conceptuelles (contextes orientés compréhension) et définir le terme ;
- la seconde méthode s'appuie sur des mesures d'association pour identifier, grâce au repérage de collocations, des contextes riches en connaissances linguistiques (contexte orienté usage) et comprendre l'usage du terme.

4.2.2 Premières expérimentations sur la génération des exercices en langues spécialisées

Notre objectif est de générer des exercices de vocabulaire de spécialité sous forme d'exercices à trous, construits automatiquement à partir d'un texte de spécialité fourni par l'enseignant utilisateur de la plateforme. Le fonctionnement de cet outil est le suivant : la plateforme dispose d'un vocabulaire spécialisé et d'un corpus de textes dans le domaine étudié, que nous appellerons 'corpus support'. L'enseignant apporte un nouveau texte dans le même domaine, que nous appellerons 'texte de référence', à partir duquel seront créés les exercices. L'enseignant choisit les termes à étudier, soit à partir du texte, soit à partir des propositions provenant de la liste du vocabulaire spécialisé. Ces termes sont ensuite recherchés dans le texte de référence pour générer les phrases support de l'exercice, et également dans le corpus, pour fournir à l'apprenant d'autres exemples d'utilisation des termes dans d'autres contextes.

Pour mettre en place notre expérimentation, nous avons choisi le thème d'un cours de Master donné à l'Université de Franche-Comté, *English for Geographers*. Le sous-domaine particulier choisi est celui de géographie de l'eau, en langue anglaise. Les actions à mettre en place sont les suivantes :

- Constituer le corpus support intégrable à la plateforme, pour pouvoir fournir automatiquement des exemples en contexte. Ce corpus sert aussi lors des expérimentations pour vérifier l'adéquation de la liste de vocabulaire spécialisé à des textes du domaine ;
- Établir une liste de termes de spécialité intégrable à la plateforme ;
- Constituer un ensemble de textes de référence pour tester la plateforme.

Corpus support

Le corpus support est constitué de 44 textes en anglais provenant de Wikipédia, édités entre 2013 et 2015, soit un total de 199 448 mots. L'utilisation de l'encyclopédie en ligne Wikipédia, qui propose un large choix de textes spécialisés en accès libre, nous permet d'envisager une automatisation du choix des textes à l'avenir. Le corpus a été nettoyé manuellement et converti en format TXT.

Liste de vocabulaire de spécialité

Il s'agit de constituer une liste des termes d'intérêt de la langue de spécialité, sans être spécialiste du domaine, ce qui est le cas de l'enseignant, futur utilisateur de la plateforme. Cette liste constituera le vocabulaire de spécialité intégré à la plateforme.

Pour une première expérimentation nous avons choisi d'utiliser deux listes terminologiques déjà existantes : *International glossary of hydrology* (OMM 2012), qui contient 2059 termes et le lexique anglais-français du *Dictionnaire encyclopédique des sciences de l'eau* (RAMADE 1998), qui contient 1645 termes.

Ensemble de textes de référence

Les textes de référence sont les textes fournis par l'enseignant. C'est dans les phrases de ces textes que des termes de spécialité vont être choisis pour être remplacés par des blancs et présentés aux apprenants sous la forme des exercices à trous.

Il importe de noter que les phrases extraites du corpus support (qui sont différentes de celles du texte fourni par l'enseignant) serviront d'indice supplémentaire pour aider l'apprenant à deviner les termes à restituer dans le texte de référence.

L'ensemble des textes de référence est constitué de 20 textes sur la géographie de l'eau publiés entre 2009 et 2016, dont 10 textes scientifiques[86] et 10 textes journalistiques spécialisés[87]. Ces textes ont été nettoyés manuellement et convertis en format TXT.

[86] Issus du journal Water Research de l'International Water Association - IWA, et de l'organisation d'éducation environnementale Field Studies Council – FSC.
[87] Issus de la National Geographic Society, du magazine en ligne ScienceDaily, de la Royal Geographical Society et de la BBC.

Listes du vocabulaire de spécialité

Pour obtenir la liste du vocabulaire spécialisé à intégrer dans la plateforme, nous avons combiné deux listes terminologiques cités précédemment : *International glossary of hydrology* (A) et *Dictionnaire encyclopédique des sciences de l'eau* (B). Nous avons identifié les termes communs entre ces deux listes (leur intersection A ∩ B), l'objectif étant d'évaluer les différences entre les deux listes. Nous avons également constitué la liste de termes appartenant à la liste A ou B (leur union).

Nous avons projeté chacune de ces listes sur le corpus support, par un script qui permet d'identifier toutes les occurrences des termes dans les textes dans leurs formes au singulier et au pluriel. Le tableau 15 présente les résultats.

Liste	Nombre de termes	Nombre d'occurrences dans le corpus	Nombre de termes (uniques) dans le corpus	Pourcentage des termes qui apparaissent dans le corpus
A	2 059	11 268	565	27,44 %
B	1 645	19 748	543	33,01 %
A ∩ B	202	8 032	155	76,73 %
A ∪ B	3 502	21 779	952	27,18 %

Tableau 15 Projection des termes des listes sur le corpus support

Les listes A et B ont peu de termes en commun : 202 termes, ce qui constitue moins de 10 % pour A et moins de 13 % pour B. Nous avons constaté que les deux listes ne contiennent pas les mêmes classes sémantiques de termes dans les mêmes proportions : par exemple, la liste B contient plus de noms d'espèces vivantes que la liste A. Malgré ces disparités, les 202 termes communs nous ont permis de constituer une catégorie expérimentale de termes centraux de la spécialité géographie-eau pour la sélection de termes des exercices à trous. Ces 202 termes communs sont très bien représentés dans le corpus : 155 parmi eux ont des occurrences dans le corpus.

Texte	Nombre de mots	Nombre d'occurrences	Nombre de termes	Pourcentage
Textes journalistiques (10 textes)				
1	815	120	52	6,38 %
2	2 177	306	82	3,77 %
3	1 229	228	71	5,78 %
4	1 215	151	52	4,28 %
5	915	141	54	5,90 %
6	612	100	29	4,74 %
7	1 259	261	77	6,12 %
8	4 581	85	37	0,81 %
9	491	115	46	9,37 %
10	1 833	285	78	4,26 %
Textes scientifiques (10 textes)				
11	6 694	1 040	140	2,09 %
12	10 660	1 433	109	1,02 %
13	15 051	2 971	263	1,75 %
14	10 461	2 168	151	1,44 %
15	8 833	1 947	142	1,61 %
16	14 731	2 501	202	1,37 %
17	3 561	738	145	4,07 %
18	2 312	402	84	3,63 %
19	2 901	610	111	3,83 %
20	2 494	436	87	3,49 %

Tableau 16 Projection des termes de la liste A ∪ B sur le corpus de référence

Notons également que les listes A et B sont tout à fait comparables, à la fois sur le nombre de termes et sur la proportion de termes reconnus dans le corpus. De ce fait, nous avons choisi d'intégrer l'union de ces deux listes (A ∪ B) à la plateforme.

Afin d'évaluer la pertinence de la liste A ∪ B pour la création d'exercices à partir de textes du domaine de la géographie de l'eau, nous l'avons projeté sur le corpus de textes de référence. Le tableau 16 présente les résultats. La dernière colonne donne le pourcentage des termes de la liste A ∪ B qui ont des occurrences dans le texte.

Reconnaissance de phrases significatives

Par phrase significative nous entendons une phrase qui 1) inclut le terme que l'on veut faire deviner et 2) permet de deviner ce terme, grâce à des caractéristiques contextuelles reconnaissables que nous appellerons 'indices'. Nous avons procédé au repérage manuel de phrases significatives du corpus support parmi celles contenant les termes projetés, pour établir une liste des indices, dont une partie est présentée dans le Tableau 17.

Nous associons à chaque indice détecté un poids, qui exprime son degré d'informativité ou de richesse de contexte. Nous l'appellerons poids d'indice.

Dans la phrase, le poids du terme que l'on veut deviner est obtenu par l'addition des poids de tous les indices qui le concernent dans la phrase et dans le contexte proche. Les phrases significatives seront sélectionnées en fonction du terme (ou plusieurs termes) ayant le poids le plus élevé.

Caractéristiques contextuelles (indices)	Poids
Le terme est un hapax : il n'y a qu'une seule occurrence du terme dans le texte. Exemple (1) : 'hydrological models'.	+2
Indice de description ou définition dans la phrase, après le terme. Exemple (1) : 'is the branch of', 'deals with'.	+2
Présence d'autres termes projetés dans la phrase, quel que soit leur nombre. Exemple (3) : 'source, 'sources'.	+1
Présence d'autres termes projetés dans la phrase suivante, quel que soit leur nombre. Exemple (2) : 'processes'.	+0,5
Indice d'explication dans la phrase, après le terme : 'for example','that means',...	+1,5
Indice 'faible' d'extension de la phrase, après le terme (ne compter qu'une fois chacun de ces indices, quel que soit leur nombre dans la phrase). Exemple (3) : '(', 'or', 'and', ','.	+0,5
Indice 'fort' d'extension de la phrase, après le terme : 'however', 'also', 'but',...	+1
Indice d'anaphore dans la phrase, après le terme. Exemple (1) : 'which'.	+1,75
Indice d'anaphore dans la phrase, avant le terme.	-1
Indice d'anaphore dans le 1er syntagme de la phrase suivante : 'it', 'that',… Exemple (2) : 'They'.	+1
Le terme est immédiatement suivi du verbe être, à la 3ème personne de son nombre (singulier ou du pluriel) : 'is', 'are'. Exemple (2) : 'are'.	+1,5

Tableau 17 Exemples d'indices

Voici des exemples de pondérations correspondant à des indices du Tableau 17, pour évaluer la richesse des phrases ((a) phrases d'origine, (b) mêmes phrases avec pondération, et (c) calcul du poids d'indice par addition des pondérations du contexte pour le terme) :

(1) Poids d'indice fort : 9,25 pour le contexte du terme 'hydrography' :

(1.a) « *hydrography is the branch of applied sciences which deals with the measurement and description of the physical features of oceans, seas, coastal areas, lakes and rivers,...* »

(1.b) « [hydrography] 'is the branch of'(2) applied sciences 'which'(1,75) 'deals with'(2) the measurement and(0,5) 'description of'(2) the physical features of oceans, [seas](1/4), coastal [areas](1/4), [lakes](1/4) and [rivers](1/4),... ».

(1.c) [hydrography] = 2 + 1,75 + 2 + 0,5 + + 2 + 1 = 9,25

(2) Poids d'indice moyen : 4,5 pour le contexte du terme 'hydrological models' :

(2.a) « *hydrological models are simplified, conceptual representations of a part of the hydrologic cycle. They are primarily used for hydrological prediction and for understanding hydrological processes.* ».

(2.b) « [hydrological models] 'are'(1,5) simplified ','(0,5) conceptual representations of a part of the hydrologic [cycle](1). 'They'(1) are primarily used for hydrological prediction and for understanding hydrological [processes](0,5) ».

(2.c) [hydrological models] = 1,5 + 0,5 + 1 + 1 + 0,5 = 4,5

(3) Poids d'indice faible : 3 pour le contexte du terme 'river' :

(3.a) « *A river begins at a source (or more often several sources) and ends at a mouth, following a path called a course* »

(3.b) « A [river] begins at a [source](1/2) '('(0,5) 'or'(0,5) more often several [sources](1/2) 'and'(0,5) ends at a mouth ','(0,5) following a path called a course ».

(3.c) [river] = 1/2 + 0,5 + 0,5 + 1/2 + 0,5 + 0,5 = 3

En résumé, nous proposons d'appliquer aux textes — ceux du corpus support et ceux fournis par l'enseignant-usager de la plateforme — une pondération des indices comme celle décrite ci-dessus, afin de détecter 1) les phrases significatives dans le corpus support (elles doivent servir d'exemples ajoutés dans l'exercice à trous), 2) les termes à remplacer par des blancs dans le texte de référence fourni par l'enseignant-utilisateur.

4.2.3 Conclusion

Afin de développer une plateforme pour les enseignants de langues étrangères de spécialité pour la préparation de matériels didactiques, nous avons posé les bases d'un générateur automatique d'exercices à trous en langues spécialisées. Nous avons tout d'abord conceptualisé le fonctionnement d'un tel exerciseur en nous basant sur les exerciseurs qui existent déjà pour l'apprentissage des langues étrangères (non spécialisées). Nous avons sélectionné des matériels de langues spécialisées intégrables à la plateforme selon des procédures réutilisables pour d'autres langues et spécialités, et nous avons effectué une expérimentation de méthodes de sélection des termes et des phrases dans le corpus pour

produire automatiquement des exercices à trous. Nous avons mis en place des indices permettant d'exprimer et calculer la valeur informative d'une phrase, pour qu'elle constitue un contexte significatif pour la recherche d'un terme.

Afin de générer des exercices à choix multiples, nous devons considérer les suggestions des réponses qui pourraient être données à un apprenant cherchant à deviner un terme dans une phrase à trou. Pour que ces suggestions soient cohérentes, nous avons effectué une première catégorisation sémantique pour les termes du domaine de la géographie de l'eau. Un terme peut appartenir à plusieurs catégories à la fois. Quelques exemples de catégories sont présentés dans le Tableau 18.

Catégorie	Termes
Lieu	'river bed', 'river bank', 'meander', 'mouth', 'estuary', 'delta', 'cliff face', 'coastline', 'source', 'bridge', 'harbour',...
Etat	'liquid', 'humid', 'temperate', 'tropical', 'cold', 'dry', 'hot', 'warm', 'wet', 'polar',...
Phénomène	'evaporation', 'water level', 'condensation', 'flooding', 'confluence', 'melting', 'freezing', 'erosion', 'deposition', 'attrition', 'soil erosion', 'deforestation', 'flooding', 'monsoon', 'erosion', 'abrasion', 'flow down', 'storm', 'thunderstorm', 'drought',...
Concept	geography, location, water level, weather, confluence,...
Objet naturel	'water vapour', 'glacier', 'lake', 'source', 'cloud', 'coast', 'wave', 'cliff', 'biomes', 'landscapes', 'tundra', 'channel', 'desert', 'cloud', 'ecosystem', 'environment', 'fog', 'ice', 'population', 'rain', 'sea', 'river', 'snow', 'sun', 'water', 'wind', 'sediment', 'storm', 'thunderstorm',...
Objet technique	'bridge', 'harbour', 'dam', 'aqueduct', 'map',...
Vivant	'deforestation', 'biomes', 'tundra', 'ecosystem', 'environment', 'population',...

Tableau 18 Exemples de catégories sémantiques spécifiques au domaine de spécialité

Les résultats de ces premières expérimentations montrent la faisabilité de la génération d'exercices à trous à partir de listes de vocabulaire et textes de référence. Dans le futur, nous allons développer l'outil de génération d'exercices à trous t nous allons prototyper, sur la base de ce type d'exercices, une plateforme de génération automatique d'exercices de langues spécialisées et proposer d'autres matériels pour les enseignants de langues spécialisées.

5. CONCLUSION ET PERSPECTIVES

Il importe peu que les questions restent sans réponse.

Ludwig Wittgenstein

Dans ce mémoire, j'ai présenté les divers travaux que nous avons entrepris pour la conception d'outils d'aide à la rédaction des textes en langues spécialisées et leur contexte de développement. J'ai parlé de l'importance de la modélisation linguistique, des contraintes liées à l'automatisation des certaines tâches, du rôle de l'expert et de la prise en compte des utilisateurs finaux. J'ai mentionné l'importance du lexique et de son contextualisation, une thématique, qui, je pense, fait consensus parmi les chercheurs qui s'intéressent à la rédaction en langues spécialisées. J'ai présenté quatre logiciels (certains encore au stade expérimental) qui sont conçus pour donner à un auteur d'un texte spécialisé plus d'autonomie dans la rédaction. Certains de ces logiciels s'adressent directement aux rédacteurs (le Compagnon LiSe et l'outil d'aide à la rédaction médicale). D'autres constituent des aides indirectes, soit à l'enseignement et l'apprentissage du vocabulaire spécialisé du niveau académique, soit à la constitution des lexiques pour la rédaction en langues contrôlées. Ces logiciels tentent timidement de répondre au défi de la production langagière technique, en explorant des pistes que Jean-Marie Klinkenberg définissait déjà comme prioritaires en 1997, lors du colloque sur la rédaction technique à Bruxelles. Il s'agissait d'encourager l'apprentissage de la rédaction technique, d'agir sur les entreprises pour qu'elles reconnaissent l'importance du travail rédactionnel et surtout d'encourager le développement des outils d'aide à la rédaction.

Les outils d'aide à la rédaction répondent à un véritable besoin sociétal, surtout que les rédacteurs des écrits en langues spécialisés ne sont quasiment jamais des professionnels de la rédaction. Par conséquent, les logiciels que nous développons ne s'adressent pas en priorité à des rédacteurs techniques formés, mais à des rédacteurs 'occasionnels'. Le concept de rédacteur 'occasionnel' est de fait un euphémisme, puisque 60% des cadres sont appelés à rédiger des rapports ou des articles au moins une fois par semaine (REJEAN 2000). Ceci est confirmé par notre expérience de terrain : les institutions avec qui nous avons travaillé (EFS B/FC, CHRU de Besançon) n'emploient pas de rédacteurs techniques formés (ce sont souvent les cadres de santé, infirmiers et médecins qui rédigent la documentation opérationnelle, ensuite validée par les responsables qualité). Il en va de même pour la rédaction d'articles scientifiques : plus de 90% des professionnels de santé qui ont répondu à notre enquête sur leurs habitudes rédactionnelles en anglais déclarent rédiger leurs articles sans assistance. Ceux qui font appel à un traducteur (51.5%), le font surtout pour la révision d'un article déjà rédigé et seulement 14,3% ont recours à un traducteur pour traduire leur manuscrit français en anglais.

La spécificité de ces rédacteurs occasionnels – qu'il s'agisse de rédaction technique ou scientifique - impose de véritables contraintes sur la conception des outils d'aide à la rédaction, notamment à cause de la nature chronophage de cette activité qui n'est pas leur préoccupation principale. Nous avons essayé d'en tenir compte dans nos travaux, mais nous

nous trouvons encore aujourd'hui devant un certain nombre des défis à relever. Ces défis résultent d'une injonction apparemment contradictoire : d'une part, du besoin des outils simples, mais précis, et d'autre part, de la complexité de l'information à transmettre. En effet, il existe déjà d'excellents dispositifs qui pourraient aider la rédaction et qui se donnent en partie, cette finalité[88]. Cependant, ils ne sont pas encore tout à fait appropriés aux besoins d'un rédacteur technique occasionnel. À mon avis, plusieurs aspects dans la conception de ces outils doivent être approfondis:

1. **La spécialisation par le sujet** : c'est une demande récurrente de nos interlocuteurs métier, dans l'ensemble de nos projets. La spécialisation est entendue ici comme la limitation du sujet à un domaine très précis, délimité soit par le cadre institutionnel (comme c'était le cas de l'EFS B/FC), soit par le domaine, (voire sous-domaine, comme c'est le cas dans le biomédical). Au niveau recherche, cette injonction pose la problématique de la création d'outils facilement adaptables et paramétrables en fonction des sujets. On imagine volontiers la possibilité de déterminer le périmètre de son corpus par un utilisateur, la possibilité de l'élargir ou le restreindre en fonction du lexique à traiter, de choisir les types d'informations à afficher, tout en conservant la fiabilité de l'information. Il existe déjà des outils très performants qui disposent de ces fonctionnalités comme le Sketch Engine[89] (KILGARRIFF 2012) pour l'extraction des collocations. C'était aussi l'intention derrière le développement de la Station Sensunique, mais ces outils sont plutôt destinés et utilisés par les professionnels de la langue (linguistes, lexicographes, traducteurs, enseignants des langues etc.) qui acceptent une certaine complexité. Le défi consiste alors à concevoir un outil de ce type pour les rédacteurs, tout en gardant à l'esprit qu'ils ne sont pas prêts à investir de temps dans l'apprentissage.

2. **Conception d'outils orientés vers la production** (à l'instar des dictionnaires orientés vers l'encodage). Les dictionnaires d'encodage (par opposition à des dictionnaires de décodage) sont des dictionnaires destinés à aider la production langagière. Le premier dictionnaire de ce type est certainement le Dictionnaire Explicatif et Combinatoire du Français Contemporain (DEC) d'Igor MEL'CUK que son auteur définit de la façon suivante : '(DEC) is intended to supply all the information which is conveyed by individual lexical units and which is necessary to express a given meaning in a completely idiomatic way' (MEL'CUK 1988 : 167). Le DEC est un dictionnaire de la langue générale, employant un dispositif linguistique des descriptions des unités lexicales trop complexe pour un utilisateur non-initié. Son ambition est aussi de fournir toutes les informations nécessaires à l'encodage d'une unité lexicale, ce qui le rend très précis, mais aussi très difficile à utiliser. D'autres dictionnaires spécialisés, pour la plupart définis comme les dictionnaires d'apprentissage, proposent des dispositifs novateurs de description de l'information linguistique, par exemple DiCoInfo (L'HOMME 2008), dans la lignée de l'école melčukienne, DAFA (BINON et al. 1992) ou the Louvain English for Academic Purposes Dictionnary (LEAD) (GRANGER & PAQUOT 2010a, 2010b ; PAQUOT 2012), tous les deux développés à l'Université de Leuven. Dans nos travaux sur l'aide à la rédaction, nous auditionnons l'expérience de ces dictionnaires, surtout pour décider du périmètre des données à décrire et de trouver la meilleure façon de les présenter, compte

[88] Par exemple, les deux bases de données issues des projets ARTES et Scientext.
[89] https://www.sketchengine.co.uk/#blue

tenu du profil des utilisateurs. Mais nous nous interrogeons surtout sur ce que serait un véritable outil orienté vers la production de contenu pour lequel il faudra distinguer entre les informations déjà connues de l'utilisateur et celles que l'outil devrait lui apporter, soit de façon dictionnairique, soit en ayant recours à un moteur de recherche. Nous poursuivons ses travaux dans le projet SARS, mais aussi dans deux autres projets que nous sommes en train de finaliser : le projet Trad'Arc et le projet MeNuSA.

5.1 Projet TradArc

Le projet TradArc[90] réunit le Centre L. Tesnière (coordination : Izabella Thomas) et le Département de Traitement Informatique Multilingue de la Faculté de Traduction et d'Interprétation de l'Université de Genève (coordination : Pierrette Bouillon) et deux start-ups spécialisées dans les technologies langagières, situées des deux côtés de la frontière (translat.me, France et Modulo Language, Suisse). Ce projet a deux objectifs principaux :

- Mettre en place des outils et des ressources pour la rédaction scientifique et technique, ainsi que la traduction spécialisée, afin d'améliorer les contenus linguistiques des entreprises, dans deux de domaines d'activité phares de l'Arc Jurassien : le biomédical et le patrimoine culturel ;
- Centraliser ces ressources pour en faciliter l'accès, en vue de mieux accompagner linguistiquement le déploiement international de services et de produits de pointe issus de ce territoire, aussi bien grâce aux technologies de la rédaction et de la traduction qu'aux compétences humaines (annuaire de traducteurs et jeunes traducteurs spécialistes de ces domaines).

Dans ce but, on propose de structurer la chaîne de production et de traduction de la documentation en trois étapes et de les doter chacune d'outils d'assistance spécifiques, sous différentes formes :

- Aide à la rédaction technique et scientifique pour le français : glossaires terminologiques, logiciel d'aide à la rédaction technique, logiciel de pré-édition;
- Traduction automatique (TA) et post-édition évoluée : systèmes de traduction automatique statistiques vers l'anglais et l'allemand ; système de post-édition évoluée, adaptée aux domaines choisis ;
- Révision et contrôle qualité : modules de contrôle qualité automatique et manuel.

Certains des modules qui seront développés dans ce projet se fondent sur des technologies existantes, comme la TA, mais exigent un travail important d'adaptation aux domaines pour fournir des résultats exploitables. D'autres seront développées dans le cadre du projet, comme les outils d'aide à la rédaction. Tous les outils proposés seront évalués, puis rendus accessibles sous forme d'API[91] pour faciliter leur intégration dans les systèmes informatiques des industries partenaires.

[90] En cours de soumission au programme Interreg France-Suisse 2014 – 2020, le projet devrait commencer en septembre 2017.
[91] Une API (Application Programming Interface) ou interface de programmation est une interface par laquelle un logiciel offre des services à d'autres logiciels, souvent par le biais du web. Elle permet d'intégrer les fonctionnalités d'un logiciel (par ex. un service de traduction automatique) à un autre logiciel (par ex. une plateforme web de traduction collaborative).

5.2 Projet MeNuSA

Le projet MeNuSA (Médecine Nucléaire Sans Ambiguïté)[92] est un projet de recherche collaboratif porté par les Services hospitaliers du CHRU à Besançon et du CHUV à Lausanne. Son objectif principal est d'étudier la communication sur la radioprotection en Médecine Nucléaire afin de fournir *in fine* :

- aux patients et à leurs aidants : des informations claires, faciles à comprendre et non anxiogènes ;
- aux professionnels de santé chargés de la production de ces informations : des moyens de communiquer plus facilement et efficacement.

Les observations dans les Services de Médecine Nucléaire et les conclusions de projets antérieurs, associant chercheurs en santé et en linguistique et focalisés sur les professionnels de santé spécialisés, suggèrent que les consignes de radioprotection, transmises aux patients et/ou à leurs aidants, avant ou après un examen médical irradiant, peuvent être mal comprises et peuvent provoquer des réactions émotionnelles négatives. À cela s'ajoute le fait que dans l'imaginaire du grand public le mot « nucléaire » suscite très souvent aujourd'hui méfiance et inquiétude. Ces réactions (parfois inappropriées et délétères) concernent au premier chef les patients, mais aussi leurs aidants, qu'il s'agisse d'aidants familiaux ou amicaux bénévoles, ou des personnels à bas niveau de formation à domicile et dans les institutions extrahospitalières de long séjour, tout particulièrement pour les patients âgés.

En plus des spécialistes en Médecine Nucléaire, le projet associe les chercheurs en Sociologie, en Anthropologie et en Ethnologie (Laboratoire de Sociologie et Anthropologie, LASA de l'Université de Franche-Comté – UFC / Unité de recherche en santé de la Haute Ecole de Santé Vaud - HESAV), en linguistique et en communication technique (Centre L. Tesnière de l'UFC, Prolipsia / Université de Lausanne - UNIL) et en communication et santé (Cévéa Santé).

Ainsi, pour atteindre l'objectif principal du projet, les tâches/étapes suivantes ont été identifiées :

- Étudier la chaine de transmission des informations liées à la radioprotection dans sa double composante orale et écrite afin d'objectiver leur cohérence et leur effet sur les patients et leurs aidants ;
- Analyser les phénomènes comportementaux indésirables associés aux informations diffusées aux patients et aux aidants, susceptibles de nuire à la prise en charge globale des pathologies concernées ;
- Auditer les documents mis à disposition des patients et aidants afin d'évaluer leur niveau de compréhension des textes de radioprotection, tout en les replaçant dans leur contexte socio-anthropologique pour proposer des solutions d'amélioration ;

[92] Ce projet est en cours de construction pour être soumis au programme Interreg 2014-2020.

- Proposer aux professionnels de soins des solutions (logiciel d'aide à la rédaction, par exemple) pour consolider le processus de production et de diffusion des informations destinées aux patients et aux aidants ;
- Développer un modèle de communication performant et facile à utiliser, et des outils variés susceptibles d'être déclinés dans d'autres domaines médicaux (autres types d'explorations complémentaires, comme les interventions ambulatoires, par exemple) ;
- Apporter aux patients et à leurs aidants des informations claires, faciles à comprendre et non anxiogènes, pouvant être intégrées dans une application pour smartphone dans le cadre des programmes d'information proposés par des compagnies d'assurance.

5.3 Et à long terme…

Le projet MeNuSA introduit bien la thématique de recherche qui m'intéresse à développer à plus long terme. Elle concerne la simplification de la communication et surtout de l'écriture institutionnelle envers les usagers.

Cette thématique, peu traitée en France, est, par exemple, une véritable préoccupation des chercheurs et de pouvoirs publics au Canada[93]. Le concept - clé de cette démarche est celui de *littératie* défini comme « *la capacité d'un individu de capter l'information (orale, écrite, graphique, gestuelle, tactile, olfactive), de la traiter et d'agir selon son bagage et les facteurs qui conditionnent cette capacité dans un domaine en particulier* » (COLLETTE & al. 2012). Or, le Conseil Canadien sur l'apprentissage rapporte que plus de 48% des adultes canadiens (personnes âgées de plus de 16 ans) souffrent d'un faible niveau de littératie, synonyme de difficultés à lire, à écrire et à bien comprendre l'information écrite mise à leur disposition[94]. La recherche pour la clarification du langage, que ce soit dans le domaine juridique (WAGNER et CACCIAGUIDI-FAHY (eds.) 2006), administratif ou en santé (RICHARD & LUSSIER 2009) est accompagnée de nombreux outils d'aide à la rédaction sous forme de manuels de rédaction[95], de testeurs du niveau de la difficulté d'un texte[96] ou de lexiques simplifiés[97]. Mais il n'existe pas d'outils automatisés à l'écriture de textes simplifiés pour le français et c'est évidemment cet aspect que je souhaiterais développer, d'autant plus qu'il existe un lien très fort entre l'écriture simplifié et l'écriture en langues contrôlées. En plus de la simplification, l'aspect le plus saillant relève de la démarche prescriptive à mettre en place pour écrire en langues simplifiées : un rédacteur en langue simplifiée se trouve devant le défi d'un vocabulaire et d'une syntaxe restreints. Il doit, alors, développer une capacité de

[93] Voir Technostyle, 2008, Numero 1, volume 22, journal de l'Association canadienne des professeurs de rédaction technique et scientifique, édité grâce à l'appui du Centre d'écriture de University of the Fraser Valley, Abbotsford, BC.
Mais c'est aussi une préoccupation en Belgique, aux pays scandinaves et autres pays anglophones.
[94] « L'avenir de la littératie dans les métropoles canadiennes », rapport de Conseil canadien sur l'apprentissage (CCA), en ligne : http://www.bdaa.ca/biblio/apprenti/cca/futureliteracy2010_fr/futureliteracy2010_fr.pdf [accédé le 09/10/2016]
[95] A noter cette initiative française : le Guide pratique de la rédaction administrative (2002), réalisé par le Centre de Linguistique Appliquée(CLA) de Besançon à la demande du Ministère de la Fonction Publique.
[96] Par exemple Scolarius, http://www.scolarius.com/; mais aussi Doxilog, un outil développé au Centre Tesnière par Yves Bordet (mais dans un objectif différent), http://www.doxilog.com/
[97] Par exemple, Le lexique des termes administratifs, http://www.tradulex.com/Seattle2005/LEXIQUE.pdf.

paraphrase pour exprimer le plus exactement le contenu des écrits juridiques, administratifs ou médicaux.

La mise en place des langages simplifiés dans les divers domaines de l'écrit institutionnel pourra faire l'objet de mes recherches dans les années à venir.

6. REFERENCES 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].

Allen J., Barthe K. (2003). Controlled Translation: the integration of Controlled Language (CL) and Machine Translation (MT), EAMT/CLAW 2003 (7th international Workshop of the European Association for Machine Tranlation and 4th Controlled Language Applications Workshop).

Almqvist I., Hein A.S. (1996). Defining ScaniaSwedish - a Controlled Language for Truck Maintenance. In Proceedings of the 1st Int. Workshop on Controlled Language Applications, CLAW'96, 159-164, KU Leuven, Belgium.

Aluísio S. M., Barcelos I., Sampaio J., & Oliveira O. N. (2001). How to learn the many unwritten" rules of the game" of the academic discourse: a hybrid approach based on critiques and cases to support scientific writing. In *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on* (pp. 257-260). IEEE.

Antoniadis G., Echinard S., Kraif O., Lebarbe T., Ponton C. (2005). Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO. In *Alsic*, vol. 8, n° 2 spécial Atala.

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.

Baker M. (1988). Sub-Technical Vocabulary and the ESP Teacher: An Analysis of Some Rhetorical Items in Medical Journal Articles. Reading in a Foreign Language, 4(2), 91–105.

Barcellini F., Grosse C., Janier M., Kang J., Torralla M. S. P., Quatrain Y. & Saint-Dizier P. (2014), Lelie: analyse et prévention des risques à travers l'aide à la rédaction de documents techniques. Congrès Lambda MU 19 de l'Institut pour la Maîtrise des Risques.

Barnes C., Boutron, I., Giraudeau B., Porcher R., Altman D. G., & Ravaud, P. (2015). Impact of an online writing aid tool for writing a randomized trial report: the COBWEB (Consort-based WEB tool) randomized controlled trial. *BMC medicine*, *13*(1), 1.

Barthe K. (2004). ASD (AECMA) Simplified English, Extracts from the Airbus France Training Course for Authors.

Beddar M. (2013). *Vers un prototype de traduction automatique contrôlée français/arabe appliquée aux domaines à sécurité critique*, Thèse de doctorat sous la direction de Sylviane Cardey, Centre Tesnière, Université de Franche-Comté, Besançon.

Bertin M., Atanassova I., Larivière V., and Gingras Y. (2015). The Invariant Distribution of References in Scientific Papers. *Journal of the Association for Information Science and Technology (JASIST)*, doi: 10.1002/asi.23367.

Binon J. & Verlinde S. (1992). Le dictionnaire d'apprentissage du français des affaires. In *Euralex '92 Proceedings I-II. The 5th Euralex international Congress on Lexicography in Tampere* (pp. 43-50).

Bouffier A. (2006). Segmentation et structuration de textes procéduraux pour l'aide à la modélisation de connaissances : le rôle de la structure visuelle. Schadae, prépublication n°10, fascicule n°1, pp. 79-84.

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).

Bourigault D. (1994). Extraction et structuration automatique de terminologie pour l'aide à l'acquisition des connaissances à partir de textes. Actes du 9ème congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'94), 1994, p. 397-408.

Bourigault D. (1994). Lexter: Un logiciel d'extraction de terminologie: Application à l'acquisition des connaissances à partir de textes. EHESS, Paris.

Bourigault D., Jacquemin C. (2000). Construction de ressources terminologiques. In J.-M. Pierrel (éd.) *Industrie des langues*. Hermès, Paris, 2000, pp. 215-233.

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.

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.

Cabré M. T. (1998). La terminologie : théorie, méthode et applications. Armand Colin ; Presses de l'Université d'Ottawa.

Cabré M.T., Estopà R., and Vivaldi J. (2001). Automatic term detection: A review of current systems. In *Bourigault D., Jacquemin C. Et L'Homme M.-C. (eds.) Recent Advances in Computational Terminology*. Amsterdam / Philadelphie, John Benjamins, pp. 53–87.

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

Campion M. E. & Elley W. B. (1971). An academic vocabulary list. Wellington, N.Z.: New Zealand Council for Educational Research.

Cardey S. (2013). *Modelling Language (Natural Language Processing)*. John Benjamins Publishing Company.

Cardey S. (2009). Controlled Languages for More Reliable Human Communication in Safety Critical Domains. In the 11th International Symposium on Social Communication, ACTAS, Santiago de Cuba, 2009, pp 330-335.

Cardey S., Greenfield P., Vienney S. (2005). Machine Translation, Controlled Languages and Specialised Languages Lingvisticae ligature Investigationes, Benjamins.

Cerbah F., Daille B. (2006). Une architecture de services pour mieux spécialiser les processus d'acquisition terminologique. Traitement Automatique des Langues, vol. 47, n° 3, 2006, p. 39-61.

Charnock Ross (1999). Les langues de spécialité et le langage technique : considérations didactiques, *ASP,* n° *23-26*, 281-302.

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.

Chung M. (2009). The newspaper word list: A specialised vocabulary for reading newspapers. JALT Journal, 31(2), 159–182.

Church K. W., Hanks P. (1990). Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, vol. 16, 1990, p. 22-29.

Collette K., Rousseau J., Clerc I. et Clamageran S. (2012).Littératie et droits en matière de santé et de services sociaux. *Communication* [En ligne], Vol. 30/1 | 2012, mis en ligne le 29 novembre 2012, consulté le 09 octobre 2016. URL : http://communication.revues.org/2939 ; DOI : 10.4000/communication.2939

Coombe C. A. (2011). Assessing Vocabulary in the Language Classroom. In Anderson & Sheehan (Eds) Focus on Vocabulary: Emerging Theory and Practice for Adult Language Learners. 111-124. HCT Press, Abu Dhabi.

Coxhead A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*(2), 213–238.

Daille B. (1994) Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques, Thèse de Doctorat en Informatique Fondamentale, Université Paris 7, 1994.

Daille B. et al. (1994). Towards automatic extraction of monolingual and bilingual terminology. Proceedings of the 15th conference on Computational linguistics, 1994, p. 515-521.

Depecker L. & Roche C. (2007). Entre idée et concept: vers l'ontologie. *Langages*, (4), 106-114.

Drouin P. (2002). Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés, Thèse de Doctorat en Linguistique, Université de Montréal, 2002.

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.

Drouin P. (2007). Identification automatique du lexique scientifique transdisciplinaire. Revue française de linguistique appliquée, vol. 12, no 2, p. 45-64

Dubay W.H. (2004) The principles of readability, Impact Information, 2004. http://www.impact-information.com/impactinfo/readability02.pdf

Dunning T.E. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, vol. 19/1, 1993, p. 61-74.

Ecarnot F., Seronde M. F., Chopard R., Schiele F., & Meneveau N. (2015). Writing a scientific article: A step-by-step guide for beginners. *European Geriatric Medicine*, *6*(6), 573-579.

Enguehard A. (1992). ANA, Apprentissage Naturel Automatique d'un Réseau Sémantique, Thèse de Doctorat en Science Spécialité Contrôle des Systèmes, Université de Technologie de Compiègne, 1992.

Evert S., Heid U., & Spranger K. (2004). Identifying Morphosyntactic Preferences in Collocations. In *LREC*.

Falaise A., Tutin A., & Kraif O. (2012). Une interface pour l'exploitation de corpus arborés par des non-informaticiens: la plate-forme SCIENQUEST du projet Scientext. *Traitement Automatique des Langues*, *52*(3), 201-228.

Fiscus J.G. (1997). A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER). *In IEEE Workshop on Automatic Speech Recognizer and Understanding,* p. 347-354.

Fraser S. (2007). Providing ESP Learners with the Vocabulary They Need : Corpora and the Creation of Specialized Word Lists. Hiroshima Studies in Language and Language Education, Issue 12, 127–143.

Fraser S. (2009a). Technical vocabulary and collocational behaviour in a specialised corpus (pp. 3–5).

Fraser S. (2009b). Breaking Down the Divisions between General, Academic, and Technical Vocabulary : The Establishment of a Single, Discipline-based Word List for ESP Learners. Hiroshima Studies in Language and Language Education, (12), 151–167.

Gavieiro-Villatte E., Spaggiari L. (1999). Open-ended overview of controlled languages. In: BULAG N° 24, ISSN: 07958 6787, ISBN: 2-913322-62-2, pp. 89-100., 1999

Gentilhomme Y. (1995). Contribution à une réflexion sur les locutions mathématiques. *Cahiers du français contemporain*, *2*, 197-242.

Gentilhomme Y. (2000). Du sens à la définition en paysage mathématique. *Le Sens en terminologie. Lyon: Presses Universitaires de Lyon*, 218-255.

Ghadessy P. (1979). Frequency counts, word lists, material preparation: A new approach. English Teachung Forum, (17).

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

Grabar N. (2004). Terminologie médicale et morphologie : Acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique, Thèse de Doctorat en Informatique Médicale, Université Paris 6.

Granger S. & M. Paquot (2010a). Customising a general EAP dictionary to meet learner needs. In Granger S. & M. Paquot (eds) (2010) eLexicography in the 21st century: New challenges, new applications. Proceedings of ELEX2009. Cahiers du CENTAL. Louvain-la-Neuve, Presses universitaires de Louvain, 87-96.

Granger, S. & M. Paquot (2010b). The Louvain EAP Dictionary (LEAD). In Proceedings of the XIV EURALEX International Congress, Leeuwarden, The Netherlands, 6-10 July 2010, 321-326.

Guide pratique de la rédaction administrative (2002). Paris, Ministère de la Fonction publique et de la réforme de l'État, réalisé par le Centre de Linguistique Appliquée de Besançon sous l'autorité du Comité d'orientation pour la simplification du langage administratif, 111 p.

Gurevych I., Bernhard D. et Burchardt A. (2009). Tutorial Notes - Educational Natural Language Processing. *AIED 2009*, Brighton. Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt, Allemagne.

Hamon T. (2000). Variation sémantique en corpus spécialisé : Acquisition de relations de synonymie à partir de ressources lexicales, Thèse de Doctorat en Informatique, Université Paris Nord.

Heurley L. (2001/2002). Compréhension et utilisation de textes procéduraux : l'effet de l'ordre des informations. Revue Française de Linguistique Appliquée, Volume VI, pp 29-46.

Hmida F., Morin E. et Daille B. (2015). Extraction de Contextes Riches en Connaissances en corpus spécialisés. In Actes de la *22ème conférence sur le Traitement Automatique des Langues Naturelles*, 425-431, Caen.

Hoffmann Lothar. Toward a Theory of LSP. *Fachsprache. Internazionale Zeitschrift für Fachsprachenforschung,-didaktik und Terminologie Wien*, 1979, vol. 1, no 1, p. 12-16.

Hsu W. & al. (2011). A business word list for prospective EFL business postgraduates. Source: The Asian ESP Journal, 7(4), 63–99.

Huguier M. & Maisonneuve H. (2003). *La rédaction médicale*. Wolters Kluwer France.

Humbley J. (2004). La réception de l'œuvre d'Eugen Wüster dans les pays de langue française. Les Cahiers du CIEL, p. 33- 51.

JIN G. (2015). *Système de traduction automatique français–chinois dans le domaine de la sécurité globale*, Thèse de doctorat sous la direction de Sylviane Cardey, Centre Tesnière, Université de Franche-Comté, Besançon.

Kilgarriff A., Rychlý P., Kovář V. and Baisa V. (2012). Finding Multiwords of More Than Two Words. In *Proceedings of the 15th EURALEX International Congress*, Norway, pp. 693–700.

Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, *6*(1), 97-133.

Kinnunen T., Leisma H., Machunik M., Kakkonen T., & Lebrun J. L. (2012). SWAN-scientific writing AssistaNt: a tool for helping scholars to write reader-friendly manuscripts. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 20-24). Association for Computational Linguistics.

Klinkenberg, J. M. (2000). Introduction à la problématique. In : B. Denis (eds.) La rédaction technique. Coll. *Champs linguistiques*, De Boeck Supérieur, 11-24.

Klinkenberg, J. M. (2002). La légitimation de la variation linguistique. *L'Information grammaticale*, *94*(1), 22-26.

Kübler N. & Pecman M. (2012). The ARTE bilingual LSP dictionary: From collocation to higher order phraseology. In *Electronic Lexicography*, S. Granger & M. Paquot (eds). Oxford: Oxford University Press, pp 187-209.

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., Robichaud B. & Leroyer P. (2013). Encoding collocations in DiCoInfo: From formal to user-friendly representations. In Electronic Lexicography (pp. 211–236). Oxford: Oxford University Press.

L'Homme M.-C. (2004). La terminologie: principes et techniques. Montréal, Canada, Presses de l'Université de Montréal, 2004.

L'Homme M.-C. (2008). Le DiCoInfo: Méthodologie pour une nouvelle génération de dictionnaires spécialisés. Traduire, 217, 78–103.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, *1*(1), 127-165.

Lam J. K. (2001). A study of semi-technical vocabulary in computer science texts, with special reference to ESP teaching and lexicography.

Laroche L. (2012). Méthodologie d'établissement d'une langue contrôlée : application d'une langue contrôlée généralisante à un domaine spécifique, mémoire de Master 2ème année, spécialité Sciences du langage, option Traitement Automatique des Langues, sous la direction de Izabella Thomas, UFC.

Lebrun J. L. (2011). *Scientific writing 2.0: a reader and writer's guide*. World Scientific.

Lefevre M., Guin N., Cablée B. et Buffa B. (2015). ASKER : un outil auteur pour la création d'exercices d'auto-évaluation. Atelier Evaluation des Apprentissages et Environnements Informatiques – EAEI, *Conférence EIAH 2015*, Agadir, Maroc.

Lerat P. (1995). Langues spécialisées. PUF, Paris.

Leseigneur D. (1999). GIFAS Rationalized French: A Controlled Language for Aerospace Documentation in French, Technical Communication.

Lessard-Clouston, M. (2006). Breadth and depth specialized vocabulary learning in theology among native and non-native English speakers. Canadian Modern Language Review, 63(2), 175–198.

Lynn R. W. (1973). Preparing Word-Lists: A Suggested Method. RELC Journal.

Malafeev Alexey Yurievich, (2015). Exercise Maker: Automatic Language Exercise Generation. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2015)* n° 14(21), 441-452. Russian State University for the Humanitie. National Research University Higher School of Economics, Nizhny Novgorod, Russie.

Martinez I. A., Beck S. C., & Panza C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. English for Specific Purposes, 28(3), 183–198.

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.

Mel'čuk, I. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, *1*(3), 165-188.

Meyer I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In B. DIDIER, J. CHRISTIAN et M.-C. L'HOMME (Eds), *Recent Advances in Computational Terminology*, 279–302. Cité par : (HMIDA et al., 2015).

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.

Moseley Jack E. (1991). *Publier en anglais: manuel destiné aux auteurs scientifiques de langue française*. Paris, France: Éd. Tempo médical.

Mudraya O. (2006). Engineering English: A lexical frequency instructional model. English for Specific Purposes, 25(2), 235–256.

Muegge U. « Controlled language - Does my company need it? », tcworld, vol. 2, http://works.bepress.com/uwe_muegge/6, p. 16-19.

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.

O'Brian, Sharon. An Analysis of Several Controlled English Rule Sets. Presented at EAMT/CLAW2003, Dublin, Ireland, 15-17 May 2003, http://www.ctts.dcu.ie/presentations.html

Oliveira Jr O. N., Zucolotto V., & Aluísio S. M. (2006). Developing strategies to produce better scientific papers: a Recipe for non-native users of English. *arXiv preprint cs/0611013*.

OMM (2012) Organisation météorologique mondiale - OMM et Organisation des Nations unies pour l'éducation, la science et la culture - UNESCO. « Glossaire international d'hydrologie », publication OMM n° 385, 3ème édition (ouvrage quadrilingue anglais, français, russe, espagnol). Genève, Suisse : OMM. URL : www.wmo.int/pages/prog/hwrp/publications/international_glossary/385_IGH_2012.pdf.

Paquot M. (2012). The LEAD dictionary-cum-writing aid: an integrated dictionary and corpus tool. In Granger, S. & Paquot, M. (eds) Electronic lexicography. Oxford University Press, 163-185.

Pecman M. & Kübler N. (2011). ARTES: an online lexical database for research and teaching in specialized translation and communication. In *Proceedings of the First International Workshop on Lexical Resources*. Ljubljana, Slovenia.

Perez-Beltrachini L., Gardent C. et Kruszewski G. (2012). Generating Grammar Exercises. In : The 7th Workshop on Innovative Use of NLP for Building Educational Applications, *NAACL- HLT Worskhop 2012*, 147-157, Montréal.

Pho Van-Minh (2015). Génération automatique de questionnaires à choix multiples pédagogiques : évaluation de l'homogénéité des options. Thèse de doctorat, LIMSI-CNRS, Université Paris Sud - Paris XI.

Picht H. & Draskau J. (1985). *Terminology: an introduction* (Vol. 2). University of Surrey, Department of Linguistic and International Studies.

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

Power R., Scott D.,Hartley A. (2003), Multilingual generation of controlled languages, in EAMT/CLAW-03, http://www.itri.brighton.ac.uk/~Richard.Power/claw03.ps

Poudat C. (2011). Le genre textuel : unité globale, unités locales. Application aux textes scientifiques. In 9èmes journées scientifiques du réseau Lexicologie, Terminologie, Traduction, 15-16 septembre 2011, Université de Paris 13.

Poudat C. (2009). Capturing the generic structure of French linguistic articles with a focus on the core features of the genre. In Corpus Linguistics Conference 2009, Liverpool, 20-23 juillet 2009.

Praninskas J. (1972). American university word list. Longman Group Limited.

Quemada B. (1978). Technique et langage. *Histoire des techniques*, Paris, Gallimard, 1146-1240.

Ramade F. (1998). Lexique anglais-français. In *Dictionnaire encyclopédique des sciences de l'eau : biogéochimie et écologie des eaux continentales et littorales*, 715-735. Paris : Édiscience International.

Rayson P., Garside R. (2000). Comparing Corpora Using Frequency Profiling. Kilgarriff, Adam & Berber Sardinha, Tony (eds) Proceedings of the Workshop on Comparing Corpora, Hong Kong, Association for Computational Linguistics, 2000, p. 1-6.

Rayson P., Leech G. N. & Hodges M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. International Journal of Corpus Linguistics, 2(1), 133-152.

Réjean R. (2000) Les outils d'aide à la rédaction : une solution aux besoins francophones en matière de rédaction ? in Benoît Denis, La rédaction technique, De Boeck Supérieur « Champs linguistiques » p. 25-54.

Renahy J, Vuitton Da, Rath B, Thomas I, De Grivel V, Cardey S.(2015). Controlled Language and Information on Vaccines: Application to Package Inserts. Current Drug Safety, Volume 10, Number 1, Bentham Science Publishers, pp. 41-48(8).

Renahy J. (2010). *Conception d'une langue contrôlée généralisante (Application aux domaines de la santé et sécurité civile)*, Thèse de doctorat sous la direction de Sylviane Cardey, Centre Tesnière, Université de Franche-Comté, Besançon.

Renahy J., Thomas I. (2009). Compagnon LiSe: A Collaborative Controlled Language Writing Assistant. In ISMTCL Proceedings, International Review BULAG, PUFC, 2009, p. 223-230.

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.

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.

Rey F.-C., Thomas I., Atanassova I. (2016a). Doter les enseignants de langues de spécialité en outils informatisés : exemple de génération automatique d'exercices basés sur le corpus, accepté à la conférence LOSP (Langues sur objectifs spécifiques : perspectives croisées entre linguistique et didactique), novembre 2016, Grenoble, France.

Rey F.-C., Thomas I., Atanassova I. (2016b), Génération d'exercices d'apprentissage de langue de spécialité par l'exploration du corpus, Atelier Enseignement des langues et TAL (ELTAL), Conférence JEP-TALN-RECITAL, Paris, 4 juillet 2016

Richard C., & Lussier M. T. (2009). La littératie en santé, une compétence en mal de traitement. *Pédagogie médicale*, *10*(3), 123-130.

Rinck Fanny (2010). L'analyse linguistique des enjeux de connaissance dans le discours scientifique. Un état des lieux. Revue d'anthropologie des connaissances 3/2010 (Vol 4, n° 3), p. 427-450.

Rinsche A, Portera-Zanotti N. (2009). Study Report to the Directorate for Translation of the European Commission, Study on the size of the Language industry in the EU, 17th August 2009.

Riot S., Guin N., Jean-Daubias S. (2004). Assistance à l'enseignant dans le cadre de l'EIAH AMBRE : conception d'un générateur de problèmes. Rapport de recherche LIRIS (stage de DEA Informatique et PFE INSA), LIRIS - CNRS.

Roche M., Heitz T., Matte-Tailliez O., & Kodratoff Y. (2004). EXIT: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In *Proceedings of JADT* (Vol. 4, pp. 946-956).

Rondeau G. (1983). Introduction à la terminologie. Deuxième édition Gaëtan Morin éditeur.

Sager J. C. (1990). *A practical course in terminology processing*. John Benjamins Publishing, Amsterdam/ Philadelphia.

Saint-Dizier P. (2012). Facets of a Discourse Analysis of Safety Requirements. NLDB12, Groningen.

Salmi Louis-Rachid and Roger Salamon. *Lecture critique et communication médicale scientifique: comment lire, présenter, rédiger et publier une étude clinique ou épidémiologique*. Issy-les-Moulineaux, France: Elsevier-Masson, DL 2012, 2012.

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).

Selva T. (2002). Génération automatique d'exercices contextuels de vocabulaire. In *Actes de TALN 2002*, 185-194, Nancy.

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.

T. Turunen (2013). Introduction to Scientific Writing Assistant (SWAN) - Tool for Evaluating the Quality of Scientific Manuscripts, M.Sc. thesis, School of Computing, University of Eastern Finland.

Tano Marcelo (2011). L'utilisation de plateformes en ligne dans l'enseignement apprentissage de l'Espagnol pour Objectifs Spécifiques. In : Innovations didactiques dans l'enseignement apprentissage de l'espagnol de spécialité grâce aux ressources technologiques, *Les cahiers du GERES (Groupe d'Étude et de Recherche en Espagnol de Spécialité), nº 4*, 77-102. Montpellier.

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).

Thomas I., Laroche L., Plaisantin-Alecu B., Betbeder M.-L., Deilles R., Renahy J., Blagosklonov O., Vuitton DA. (2015), *Computerization of a 'Controlled Language' to Write Medical Standard Operating Procedures (SOPs)*, Procedia Computer Science, Elsevier, Volume 64, pp. 95-102, ISSN 1877-0509

Thomas I., Plaisantin Alecu B., Germain B., and Betbeder M.-L. (2014a). Station Sensunique: Architecture générale d'une plateforme web paramétrable, modulaire et évolutive d'acquisition assistée de ressources, in Abel A. Et al. (eds.). Proceedings of the XVI EURALEX International Congress: The User in Focus. Bolzano/Bozen: EURAC research, Volume: II, pp.707-726 .

Thomas I., Atanassova I. (2015). *Towards the enrichment of terminological resources by scientific corpora analysis*. eLex 2015 conference: Electronic lexicography in the 21th century: Linking lexical data in the digital age. pp. 136-151. United Kingdom, August 2015

Thomas, I., Plaisantin Alecu B., Germain B., Betbeder M.-L. (2014b), La Station Sensunique, une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non terminologiques (orientée Langues Contrôlées) , in Abel Andrea, Vettori Chiara, Ralli Natascia (eds.), Proceedings of the XVI EURALEX International Congress: The User in Focus. 15-19 July 2014, Bolzano/Bozen: EURAC research, Volume: II, ISBN: 978-88-88906-84-3, pp.727-736.

Tutin A. (2007). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, *12*(2), 5-14.

Tutin, A. (2014). *L'écrit scientifique: du lexique au discours*. F. Grossmann (Ed.). Presses universitaires de Rennes.

Tutin, A., & Grossmann, F. (2002). Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif. Revue française de linguistique appliquée, VII(1), 7–25.

Volanschi A. & Kübler N. (2010). Building an electronic combinatory dictionary as a writing aid tool for researchers in biology. *In Granger, S. Paquot, M (eds.). e-Lexicography in the 21st Century: New Applications, New Challenges: 343–355. Louvain-la-Neuve: Presses Universitaires de Louvain*

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 Protocols 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.

Wagner A. et Cacciaguidi-Fahy S. (eds.) (2006). Legal Language and the Search for Clarity: Practice and Tools. Peter Lang, Collection Linguistic Insights, Bern etc., 493 pp, ISBN 978-3-03911-169-5

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a Medical Academic Word List. English for Specific Purposes, 27(4), 442–458.

Wu X. (2005). *Designing a Controlled Language for English-Chineses Machine Translation of Medical Protocols*, in Proceedings of the Workshop Machine Translation, Controlled Languages and Specialised Languages, Besançon, 5-6 May 2004, Lingvisticae Investigationes, Benjamins, 2005, pp. 123-131.

Xue G. & Nation I. S. P. (1984). A university word list. Language Learning and Communication, 3(2), 215–229.

7. ANNEXES

7.1 Exemple d'un protocole contrôlé : CHRU, Nettoyage de chambres d'hospitalisation (Médecine Nucléaire)

1. Concerne : les médecins, infirmières et aides-soignantes et la personne compétente en radioprotection

2. Application :

Cette instruction récapitule les différentes mesures de radioprotection à adopter dans les chambres d'hospitalisation afin d'optimiser la radioprotection du personnel et de l'environnement. Les patients qui séjournent dans la chambre ont reçu de l'iode 131 radioactif en dose thérapeutique.

2.1. AVANT D'ENTRER DANS LA CHAMBRE PORTEZ :

- votre dosimètre passif,
- un dosimètre opérationnel
- des gants vinyl
- un tablier plombé (EN PRESENCE DU PATIENT)

2.2. DANS LA CHAMBRE :

Quand le malade est présent ne séjournez que le temps nécessaire pour effectuer votre tache, tenez vous à distance du patient (au moins 1 mètre) si il n'est pas nécessaire de l'approcher.
Le patient élimine l'iode par les urines et par la salive, les objets en contact avec ces excrétas peuvent être contaminés :

- jetez les objets à usage unique, les crachoirs, les mouchoirs et les chiffons de ménage dans la poubelle plombée de la salle de bain.
- contrôlez le plateau repas avec le débitmètre et jetez le matériel jetable dans la poubelle plombée, si la vaisselle est contaminée rincez la et jetez l'eau dans les toilettes
- Jetez vos gants dans la poubelle plombée avant de sortir de la chambre.

2.3. ENTRETIEN DE LA CHAMBRE ET ENLEVEMENT DES DECHETS :

Le ménage sera fait le vendredi, après le départ du patient

- Décontaminer avec les sprays spécifiques le combiné du téléphone, le fauteuil, la table roulante, le siège des WC, le pourtour des toilettes avec des lingettes puis jetez les lingettes dans la poubelle plombée.
- Verser dans la cuvette des WC du liquide décontaminant et laisser agir le week end
- Tirer la chasse d'eau plusieurs fois le lundi matin
- Contrôlez les draps et la taie d'oreiller, si le détecteur enregistre une valeur supérieure au bruit de fond mesuré dans le couloir, placez la pièce concernée dans un sac en nylon étiqueté avec le nom du service et la date d'enlèvement
- Descendez les sacs de déchets et ceux de matériel contaminé dans la ZEGDR et déposez les dans le chariot d'attente de tri

Le personnel du service de Médecine Nucléaire assurera le contrôle et la gestion dans la ZEGDR

Fin de document

UFC ANR Sensunique

NETTOYAGE DES CHAMBRES D'HOSPITALISATION

1 Personnel concerné

Service de Médecine Nucléaire Hospitalisation :
- Agents des Services Hospitaliers (ASH) ;
- aides-soignantes ;
- infirmières ;
- Personne Compétente en Radioprotection (PCR) ;
- médecins.

2 Application

Nettoyage des chambres d'hospitalisation.
Radioprotection des personnes.

3 Consignes de radioprotection

Avant d'entrer dans la chambre :
Porter les équipements de protection suivants :
- dosimètre passif ;
- dosimètre opérationnel ;
- gants en vinyle.

Quand le patient est dans la chambre :
Porter un tablier plombé.
Autant que possible :
Ne pas s'attarder.
Se tenir à 1 m minimum du patient.

4 Entretien quotidien

Rappel : Les objets en contact avec les excréta du patient sont potentiellement contaminés.

Jeter les objets jetables dans la poubelle plombée.
Ex. : crachoirs, mouchoirs, gobelets.
Mesurer la quantité de radioactivité émise autour du plateau-repas, au point le plus élevé, à l'aide du débitmètre.
Se référer au chapitre *7. Mesure de la quantité de radioactivité émise autour d'un objet.*

Si la quantité de radioactivité émise autour du plateau-repas est supérieure à 1,5 fois le bruit de fond :
Rincer la vaisselle non jetable à l'eau.
Verser l'eau dans la cuvette des WC.
Sinon :
Suivre la procédure hospitalière de traitement de la vaisselle.

Quand vous avez fini l'entretien de la chambre :
Jeter les gants en vinyle dans la poubelle plombée.

Confidentiel
Page 1 sur 2

UFC ANR Sensunique

5 Nettoyage hebdomadaire

Rappel : Les objets en contact avec les excréta du patient sont potentiellement contaminés.

Tous les vendredis, après le départ du patient :
Aérer la chambre durant 30 min au minimum.
Tirer les chasses d'eau 2 fois au minimum.
Après un délai de 30 min :
Nettoyer les objets suivants avec le spray décontaminant :
- combiné du téléphone ;
- mobilier ;
- matelas.

Nettoyer les équipements sanitaires suivants avec le spray décontaminant :
- siège des WC ;
- pourtour des WC ;
- lavabo ;
- pourtour du lavabo.

Verser du liquide décontaminant dans la cuvette des WC.
Laisser agir durant le weekend.

Mettre les linges suivants dans un sac :
- draps ;
- taie d'oreiller.

[...]

Éliminer les déchets.
Se référer au chapitre *6. Élimination des déchets.*

Tous les lundis matin, avant l'arrivée d'un nouveau patient :
Tirer les chasses d'eau 2 fois minimum.

6 Élimination des déchets

Quand vous avez fini le nettoyage de la chambre :
Fermer le sac poubelle de la poubelle plombée.
Mettre le sac poubelle sur le chariot « transport des déchets vers la ZEGDR ».
Conduire le chariot « transport des déchets vers la ZEGDR » à l'entrée de la ZEGDR.
Mettre le sac poubelle sur le chariot « attente de tri ».

7 Mesure de la quantité de radioactivité émise autour d'un objet

Mesurer le bruit de fond de la quantité de radioactivité naturelle émise dans le couloir, loin de toute source radioactive, à l'aide du débitmètre.
Mesurer la quantité de radioactivité émise autour de l'objet, au point le plus élevé, à l'aide du débitmètre.

Confidentiel
Page 2 sur 2

7.2 Station Sensunique : Algorithme de pondération

Configuration par défaut

Poids terminologique (PT) :

- seuil à partir duquel une UL a pour statut T → v1=3
- variation de PT si (au moins une forme fléchie de) l'UL globale est attestée par une ressource terminologique → v2=10
- variation de PT si la tête et l'expansion sont attestées par une ressource terminologique → v3=8
- variation de PT si l'UL a été extraite d'un Corpus Support → v16=5
- variation de PT si l'UL a été extraite d'un Corpus Contrastif → v17=5

Poids de Structure Lexicale (PSL) :

- variation de PSL si l'UL globale est attestée dans une ressource terminologique → v5=-10
- variation de PSL si la matrice morphosyntaxique est ou contient un verbe (hors participe) → v6=100
- variation de PSL si la matrice morphosyntaxique est ou contient un participe passé (ou présent) adjectival (Vppe ou Vppr) → v7=10
- seuil du nombre de dérivées à partir duquel on augmente le PSL : v8=3
- variation de PSL si le nombre de dérivées de l'UL dépasse le seuil précédent (v8) → v9=3
- seuil du nombre de collocations de l'UL à partir duquel on augmente le PSL → v10=3
- variation de PSL si le nombre de collocation de l'UL dépasse le seuil précédent (v11) → v12=3
- variation de PSL si l'UL est extraite par l'extracteur « Acabit » → v13=1

Poids d'UL (PUL) :

- seuil du nombre d'extracteurs à partir duquel on augmente le poids PUL : v14=1
- variation de PUL si l'on dépasse le seuil précédent (v14) → v15=1

Algorithme

PT = 0 ;
PSL = 0 ;
PUL = 0 ;

Pour chaque UL :

TESTS SUR EXTRACTEURS
PT = PT + (Base du PT par extracteur)$^{\text{nombre d'extracteurs ayant attestées cette UL}}$

TESTS SUR RESSOURCES INTERNES et EXTERNES
Boucle sur les ressources choisies par l'utilisateur
Si une ressource est terminologique
Si (une des formes fléchies de l'UL est présente dans la ressource)
PT=PT + v2 + indice de fiabilité de la ressource

PSL=PSL+v5
PUL=PUL+1
Sinon si (la tête est présente et l'expansion est présente également)
PT=PT + v3
Sinon si (la tête est présente ou l'expansion est présente)
PT=PT +v4
Sinon
Si (une des formes fléchies de l'UL est présente dans la ressource)
PUL=PUL+1

TESTS SUR MATRICE MORPHOSYNTAXIQUE
Si la matrice morphosyntaxique est ou contient un Verbe
PSL=PSL + v6
Sinon si la matrice morphosyntaxique est ou contient un participe passé ou présent adjectival (Vppe ou Vppr)
PSL=PSL + v7

TESTS SUR DERIVEES ET COLLOCATIONS
Si le nb de dérivées > v8
PSL=PSL+v9
Si le nb de collocations > v10
PSL=PSL + v11

TESTS SUR EXTRACTEURS
Si l'UL a été extraite par Acabit
PSL=PSL + v13
Si le nb d'extracteurs > v14
PUL = PUL + v15

TESTS SUR CORPUS CONTRASTIF SUPPORT
Si l'UL a été extraite d'un corpus support
PT=PT + V16
Si l'UL a été extraite d'un corpus contrastif
PT=PT + V17

CALCUL STATUT
Si PT > v1
Statut lexical=T
Sinon
Statut lexical=G

7.3 Station Sensunique : Format d'export


```
▼<ul>
  <id>55255173b25a2a0f7a00de1e</id>
  <forme_canonique>cellules tumorales</forme_canonique>
  <validation>0</validation>
 ▼<tete_exp>
    <tete>cellule</tete>
    <expansion>tumoral</expansion>
    <outil>yatea</outil>
  </tete_exp>
 ▼<tete_exp>
    <tete>cellule</tete>
    <expansion>tumoral</expansion>
    <outil>acabit</outil>
  </tete_exp>
  <forme_flechie genre="" nombre="" mode="" temps="" personne="" frequence="1">cellules tumorales</forme_flechie>
 ▼<col_incluse>
    <forme_col_incluse>cellule</forme_col_incluse>
    <id_col_incluse>55255171b25a2a0f7a00dd29</id_col_incluse>
  </col_incluse>
  <forme_lemme>cellule tumoral</forme_lemme>
 ▼<usage>
    <type_usage>préconisé</type_usage>
    <langue_usage>fr</langue_usage>
  </usage>
 ▼<cat_ms_fonct>
    <etiquette>Nom</etiquette>
    <outil>sensunique</outil>
  </cat_ms_fonct>
 ▼<col_associee>
    <forme_col_associee>Cellules isolées</forme_col_associee>
    <id_col_associee>55255170b25a2a0f7a00dc7b</id_col_associee>
  </col_associee>
 ▼<col_associee>
    <forme_col_associee>cellules issues</forme_col_associee>
    <id_col_associee>55255170b25a2a0f7a00dc93</id_col_associee>
  </col_associee>
 ▼<col_associee>
    <forme_col_associee>cellules totales</forme_col_associee>
    <id_col_associee>55255172b25a2a0f7a00dda6</id_col_associee>
  </col_associee>
 ▼<col_associee>
    <forme_col_associee>cellules viables</forme_col_associee>
    <id_col_associee>55255176b25a2a0f7a00df07</id_col_associee>
  </col_associee>
 ▼<col_associee>
    <forme_col_associee>cellules mortes</forme_col_associee>
    <id_col_associee>55255179b25a2a0f7a00dfab</id_col_associee>
  </col associee>
```

7.4 100 familles les plus fréquentes dans le lexique trans-biomédical

Rang	Mot représentatif	Rang	Mot représentatif
1	cell	30	bacterium
2	gene	31	approach
3	infect	32	Involve
4	analyze	33	pathway
5	protein	34	complex
6	indicate	35	Target
7	differentiate	36	intervene
8	sequence	37	prevalence
9	factor	38	Assay
10	virus	39	mutation
11	formation	40	regulate
12	clinic	41	independent
13	induce	42	Expose
14	function	43	Domain
15	region	44	Therapy
16	site	45	Cancer
17	select	46	Isolate
18	strain	47	Species
19	inhibit	48	immune
20	positive	49	multiple
21	process	50	React
22	score	51	Image
23	drug	52	incubate
24	area	53	Respond
25	tissue	54	Panel
26	secondary	55	Culture
27	range	56	recipient
28	cluster	57	transmit
29	structure	58	primary
59	molecular	80	section
60	membrane	81	diet
61	adult	82	Strategy
62	concentrate	83	Status
63	contribute	84	contrast
64	residual	85	environment
65	acid	86	promote
66	normal	87	error
67	dose	88	diagnose
68	component	89	biology
69	numerical	90	mediate
70	dynamic	91	density
71	conduct	92	interval
72	vector	93	mortality
73	exclude	94	metabolic
74	phase	95	chronic
75	stimulate	96	serum
76	profile	97	procedure
77	evolution	98	versus
78	extract	99	symptom
79	link	100	inject