Génération d'exercices d'apprentissage de langue de spécialité par l'exploration du corpus

François-C. REY, Izabella THOMAS, Iana ATANASSOVA
Centre L. Tesnière, Université de Franche-Comté
30 rue Mégevand 25030 Besançon, France
fcrey@edu.univ-fcomte.fr, izabella.thomas@univ-fcomte.fr,
iana.atanassova@univ-fcomte.fr

RÉSUMÉ

Il n'existe pas de travaux sur des logiciels dédiés à l'apprentissage des langues de spécialité, ni à la création de ressources pédagogiques spécifiques. Les outils informatisés concernant les langues de spécialité se limitent aux logiciels d'aide à la traduction et à des outils d'aide au recensement terminologique. Afin de doter les enseignants de langues étrangères de spécialité d'outils prenant en compte leurs besoins spécifiques nous proposons de concevoir et de développer une plateforme pour la génération de matériels pédagogiques dans le domaine des langues de spécialités. Pour répondre à cet objectif, nous avons effectué une première expérimentation concernant la génération automatique d'exercices d'apprentissage du vocabulaire spécialisé. Nous avons choisi la génération d'exercices à trous basés sur les phrases tirées d'un corpus de textes authentiques. Les résultats de ces premières expérimentations montrent la faisabilité de la génération d'exercices à trous, à partir de listes de vocabulaire et d'un corpus spécialisés intégrés à la plateforme, ainsi que de textes fournis par un enseignant.

ABSTRACT

A Corpus Based Approach to the Generation of Exercises for Language for Specific Purposes. There are neither works on software dedicated to the learning of language for specific purposes, nor works dedicated to the creation of specific educational resources. The computerized tools concerning the languages for specific purposes are limited to the computer aided translation software and to terminological census tools. In order to equip foreign language teachers with tools taking into account their specific needs, we propose to develop a platform for the generation of educational materials in the field of language for specific purposes. In this conceptual framework, we carried out a first study on the automatic generation of vocabulary lerning exercises for language for specific purposes. We chose the generation of cloze tests based on the sentences from a corpus of authentic texts. The results of these first experiments show the feasibility of the generation of cloze tests, starting from lists of vocabulary and from a corpus wich are specialized, integrated into the platform, and from texts supplied by a teacher.

MOTS-CLÉS : apprentissage des langues assisté par ordinateur - ALAO, enseignement des langues assisté par ordinateur - ELAO, langues de spécialité - LSp, génération d'exercices, exercices de vocabulaire, exercices en contexte, anglais de spécialité - ASP

KEYWORDS: Computer-Assisted Language Learning - CALL, contextual exercises, Language for Specific Purposes - LSP, English for Specific Purposes - ESP

1. Introduction

Les apprenants sont au cœur des systèmes logiciels d'apprentissage des langues assisté par ordinateur (ALAO) et d'enseignement des langues étrangères par ordinateur (ELAO). En revanche, les enseignants de langues disposent de peu d'outils répondant à leurs besoins en matière de préparation des activités d'enseignement. Pourtant, de nombreuses tâches peuvent faire objet d'aides automatisées, simplifiant ainsi le travail des enseignants, réduisant le temps de préparation de matériel, ou facilitant la création de ressources pédagogiques complexes à élaborer. Les besoins exprimés par les enseignants de langues vont au-delà de l'utilisation de logiciels génériques de mise en forme d'exercices. Un exemple d'un tel logiciel est *Hot Potatoes*[1], qui offre la possibilité de créer plusieurs types d'exercices sous forme de pages web interactives. Ce type de logiciels s'adapte à toutes les matières puisqu'il facilite uniquement la mise en forme et non pas la création des contenus des exercices. Cette dernière tâche est parmi les plus complexes lors de la préparation de matériels pédagogiques par des enseignants. L'objectif de notre travail consiste à intégrer des ressources et matériaux spécifiques utilisés en cours de langues.

La question qui nous intéresse particulièrement est celle de l'enseignement de langues de spécialité. Il pose des problématiques bien différentes de celle de l'enseignement de langues étrangères en général. Par langue de spécialité nous entendons un sous-ensemble d'une langue naturelle utilisé par un groupe de spécialistes à l'intérieur d'un domaine du savoir ou d'une discipline technique (définition inspirée de L'HOMME [L'HOMME, 1990]). Le domaine des langues de spécialité se confond au moins en partie avec celui des *langues pour (les) spécialistes d'autres disciplines* (LANSAD), et, dans une moindre mesure, avec celui du *Vocationally (and Professionally) Oriented Language Learning* (VOLL). Le niveau de langue de spécialité qui nous intéresse est celui enseigné à l'université, par exemple l'anglais de spécialité enseigné aux étudiants en Master de géographie, qui leur permet de lire et d'écrire des articles scientifiques dans leur domaine.

Il n'existe pas de travaux sur des logiciels dédiés à l'apprentissage des langues de spécialité, ni à la création de ressources pédagogiques spécifiques. Les outils informatisés existants concernant les langues de spécialité se limitent aux logiciels d'aide à la traduction et à des outils d'aide au recensement terminologique (systèmes de gestion de terminologie, extracteurs de termes, bases de données terminologiques, etc.). Par conséquent, notre objectif est de concevoir une plateforme d'aide semi-automatisée à la création de matériels pédagogiques de langues étrangères de spécialité. Cette plateforme doit :

- répondre aux besoins des enseignants (plutôt qu'à ceux des apprenants qui n'en sont pas des utilisateurs directs) ;
- prendre en compte des particularités des langues de spécialité de niveau académique ;
- être utilisable avec plusieurs langues et plusieurs spécialités.

Dans cet article nous décrivons les expérimentations que nous avons mises en place pour la création d'un premier type d'exercices intégrés à la plateforme : des exercices à trous sur le vocabulaire de spécialité. La spécialité a été choisie parmi des thèmes enseignés en cours universitaire de langue étrangère de spécialité : l'anglais de la géographie de l'eau. L'idée mise en œuvre est qu'il est possible de préparer automatiquement ou semi-automatiquement des textes 'authentiques' en langue étrangère de spécialité, choisis et fournis par un enseignant, pour en faire des exercices ou d'autres matériaux pédagogiques pertinents. Les ressources nécessaires pour générer des exercices à trous en langue de spécialité sont :

1 http://hotpot.uvic.ca

- un lexique spécialisé dans le domaine choisi, ce qui se traduit par des listes de termes, éventuellement avec des métadonnées associées aux termes (définitions, exemples, synonymes, etc.).
- Un corpus des textes dans le même domaine, qui sera analysé par la plateforme pour générer des ressources supplémentaires pour la création des exercices.
- Les textes fournis par les enseignants-usagers, qui sont destinés à être transformés en exercices.

Dans la section 2, nous présentons un état de l'art sur les méthodes et outils pour la génération d'exercices pour l'apprentissage des langues, en particulier pour l'apprentissage du vocabulaire et les langues de spécialité. Nous abordons la notion de « contextes riches » qui est au cœur de la problématique de création d'exercices à partir de textes. Dans la section 3 nous décrivons l'expérimentation autour de la génération d'exercices et les résultats. Dans la section 4 nous discutons des pistes de travail pour l'intégration des méthodes proposées dans une plateforme semi-automatique de création de matériels pédagogiques.

2. Etat de l'art

Concevoir des environnements informatiques pour *enseigner* semble un objectif nouveau par rapport à concevoir des environnements informatiques pour *apprendre,* même si le souhait a déjà été exprimé : en 2004, Stéphanie RIOT [RIOT *et al*., 2004, p. 8] voulait donner à l'enseignant le « rôle majeur » d'utilisateur principal d'un logiciel pédagogique. Plus particulièrement, dans le domaine de la génération automatique d'exercices, le besoin d'outils dédiés aux enseignants a déjà été identifié par Alex BOULTON [BOULTON, 2007, p. 43]. Il propose de promouvoir l'exploitation de corpus en apprentissage des langues par la création de « sites qui complètent des manuels pour un travail autonome ou pour permettre à l'enseignant de créer des activités pertinentes ». Van-Minh PHO [PHO, 2015] signale que les Environnements Informatiques pour l'Apprentissage Humain « doivent fournir des moyens pour assister les enseignants dans leur tâche de génération d'exercices ».
C'est pourtant un logiciel de rédaction manuelle d'exercices qui, dans les revues-conseils, apparaît comme le plus recommandé aux enseignants pour produire des exercices : il s'agit de *HotPatatoes*. Dans son article « Quels logiciels libres pour les professeurs de langues vivantes ? », Laure PESKINE [PESKINE, 2006] répartit l'ensemble des logiciels utilisés par les enseignants en huit catégories : outils bureautiques, outils internet, dictionnaires, traitement des images, traitement du son, lecteur multimédia, création de pages pour publication sur internet, et exerciseurs. Sa dernière catégorie, *exerciseurs,* ne comporte que deux logiciels : *Jclic* et *HotPotatoes* qui ne sont pas spécifiques aux exercices de langue.

2.1. Plateformes de génération 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 liée au fait que la génération automatique d'exercices est une problématique encore jeune, en pleine expansion.

Il n'existe pas de plateforme dédiée à l'apprentissage de langues de spécialité. Au niveau de l'enseignement universitaire, on trouve les plateformes de génération d'exercices MIRTO et ASKER. Le projet MIRTO [ANTONIADIS *et al*., 2005] aborde des 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[2]; 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 de spécialité, ni les publications du *Groupe d'Étude et de Recherche en Anglais de Spécialité* (GERAS) rassemblées dans la revue *Approches linguistiques des langues spécialisées* (ASP), ni les publications du *Groupe d'Etude et de Recherche en Espagnol de Spécialité* (GERES), ne font mention de logiciels dédiés à l'apprentissage des langues de spécialité du niveau universitaire, ni de logiciels dédiés à l'apprentissage des langues étrangères de spécialité. Les logiciels qui se rapprochent le plus de ces objectifs par la spécialisation de leurs étudiants sont ceux qui aident à l'apprentissage des langues générales au niveau élémentaire dans les filières d'enseignement à l'université [TANO, 2011].

2.2. 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 tirés 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 de spécialité.

Thierry SELVA [SELVA, 2002] décrit plusieurs types d'exercices d'apprentissage de la langue générale dans le cadre du projet E*nvironnement d'apprentissage lexical interactif pour apprenants du français – ALFALEX*. 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 électronique en ligne *Dictionnaire d'Apprentissage du Français Langue Étrangère ou Seconde - DAFLES*.

Pour les langues de spécialité et le langage technique, Ross CHARNOCK [CHARNOCK, 1999] propose de travailler sur des textes courts (résumés, introductions d'articles, etc.) pour tenir compte des apprenants qui n'ont pas des bases de langue étrangère encore bien établies, tout en supposant qu'ils ont des connaissances adéquates dans la discipline. Il commente 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.

2 Par exemple, elle est actuellement utilisée pour l'enseignement de l'informatique.

Selon MALAFEEV [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 Iryna GUREVYCH [GUREVYCH *et al.*, 2009, p. 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-ème mot dans le texte, par exemple n=5 ou n=8 ;
- 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, Christine COOMBE [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'ambiguïté 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.

2.3. 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 de l'exercice 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*. [HMIDA, 2015] proposent de mettre en œuvre la notion de *Contextes Riches en Connaissances* (CRC) introduite en 2001 par Ingrid MEYER [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.

3. Expérimentation sur la génération automatique des exercices en langues de spécialité

3.1. Préparation du matériel

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 et 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 199448 mots. L'utilisation de l'encyclopédie en ligne Wikipédia est due au fait qu'elle propose un large choix de textes spécialisés en accès libre, et permet d'envisager une automatisation du choix des textes dans 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, elles, d'indices de contexte supplémentaire pour aider l'apprenant à deviner les termes remplacés par des blancs dans le texte de référence.

L'ensemble de textes de référence est constitué de 20 textes sur la géographie de l'eau publiés entre 2009 et 2016, dont 10 textes scientifiques (issus du journal *Water Research* de l'*International Water Association - IWA*, et de l'organisation d'éducation environnementale *Field Studies Council - FSC*) et 10 textes journalistiques spécialisés (issus de la *National Geographic Society*, du magazine en ligne *ScienceDaily*, de la *Royal Geographical Society* et de la *BBC*). Ces textes ont été nettoyés manuellement et convertis en format TXT.

3.2. Expérimentations et résultats

Listes du vocabulaire de spécialité

Pour obtenir la liste de 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. Cette tâche a été effectuée 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 1 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 1 : Projection des termes des listes sur le corpus support

Nous constatons que 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.

Notons également que les listes *A* et *B* sont tout à fait comparables, à la fois en nombre de termes et en 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 géographie-eau, nous l'avons projeté sur le corpus de textes de référence. Le tableau 2 présente les résultats. La dernière colonne donne le pourcentage des termes de la liste A ∪ B qui ont des occurrences dans le texte.

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 2 : Projection des termes de la liste A ∪ B sur le corpus de référence

Reconnaissance de phrases significatives

Par *phrase significative* nous entendons une phrase qui, premièrement, inclut le terme que l'on veut effacer par un blanc, et, deuxièmement, 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 sont présentés dans le tableau 3.

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

Dans la phrase, le poids du terme que l'on veut effacer par un blanc est obtenu par l'addition des poids de tous les indices qui le concernent dans la phrase et dans le contexte proche. Nous proposons de choisir les phrases significatives en tenant compte du poids du terme le plus élevé (ou plusieurs termes de poids 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, quelque soit leur nombre. Exemple (3) : 'source, 'sources'.	+1
Présence d'autres termes projetés dans la phrase suivante, quelque 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 3 : Exemples de caractéristiques contextuelles

Voici des exemples de pondérations correspondant à des indices du tableau 3, 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-usager.

4. 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 de spécialité. 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 de spécialité 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 effectuée une première catégorisation sémantique pour les termes du domaine « géographie-eau ». Un terme peut appartenir à plusieurs catégories à la fois. Quelques exemples de catégories sont présentés dans le tableau 4.

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', 'confluenc'e, '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 4 : 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, et nous allons construire, sur la base de ce type d'exercices, un prototype opérationnel de la plateforme de génération automatique d'exercices de langues de spécialités et autres matériels pour les enseignants de langues de spécialité.

Références

ANTONIADIS Georges, ECHINARD S., KRAIF Olivier, LEBARBE T., PONTON Claude (2005), « Modelisation de l'integration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO », in *Alsic*, vol. 8, n° 2 spécial Atala.

BOULTON Alex (2007). « Esprit de corpus: Promouvoir l'exploitation de corpus en apprentissage des langues ». *Texte et Corpus,* n° 3, 37-46.

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

COOMBE Christine 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.

GUREVYCH Iryna, BERNHARD D. et BURCHARDT A. (2009). « Tutorial Notes - Educational Natural Language Processing », *AIED 2009*, Brighton. Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt, Allemagne.

HMIDA Firas, 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.

L'HOMME Marie-Claude (1990). « Y a-t-il une langue de spécialité ? Points de vue pratique et théorique » in revue *Langues et linguistique*, numéro spécial Journées de linguistique, 2011, 26-33. Centre international de recherche en aménagement linguistique, Québec.

LEFEVRE Marie, 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, *Conféerence EIAH 2015*, Agadir, Maroc.

MEYER Ingrid (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].

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.

Organisation météorologique mondiale - OMM et Organisation des Nations unies pour l'éducation, la science et la culture - UNESCO (2012). « 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

PEREZ-BELTRACHINI Laura, 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.

PESKINE Laure (2006). « Quels logiciels libres pour les professeurs de langues vivantes ? » in revue *Langues Modernes n° 1*. Paris : Association des Professeurs de Langues Vivantes - APLV.

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.

RAMADE François (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.

RIOT Stéphanie, 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.

SELVA Thierry (2002). « Génération automatique d'exercices contextuels de vocabulaire » in *Actes de TALN 2002*, 185-194, Nancy.

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), n° 4*, 77-102. Montpellier.