# Identification et exploitation de contextes riches en connaissances pour l'aide à la traduction terminologique

Firas Hmida

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

# Thèse de Doctorat

Firas HMIDA

*Mémoire présenté en vue de l'obtention du*
***grade de Docteur de l'Université de Nantes***
*sous le sceau de l'Université Bretagne Loire*

**École doctorale : Sciences et Technologies de l'Information, et Mathématiques**

**Discipline : Informatique, section CNU 27**
**Unité de recherche : Laboratoire des Sciences du numérique de Nantes (LS2N)**

**Soutenue le 6 février 2017**

# Identification et exploitation de contextes riches en connaissances pour l'aide à la traduction terminologique

## JURY

| | |
|---|---|
| Présidente : | **M^me^ Cécile FABRE**, Professeur des universités, Université de Toulouse II |
| Rapporteurs : | **M^me^ Cécile FABRE**, Professeur des universités, Université de Toulouse II |
| | **M. Alexandre ALLAUZEN**, Maître de conférences, Université Paris-Sud XI |
| | * * * |
| Directeur de thèse : | **M. Emmanuel MORIN**, Professeur des universités, Université de Nantes |
| Co-encadrante de thèse : | **M^me^ Béatrice DAILLE**, Professeur des universités, Université de Nantes |

# Remerciements

Je tiens en premier lieu à remercier Emmanuel Morin et Béatrice Daille d'avoir respectivement dirigé et co-encadré mon travail de recherche durant ces quatre années de thèse. Je les remercie de m'avoir honoré de travailler et d'apprendre à leur coté sur un sujet si intéressant, et d'avoir su se rendre disponibles malgré leurs engagements respectifs. Je les remercie tous les deux pour leur patience, leur pédagogie ainsi que leur qualité d'encadrement autant scientifique qu'humaine.

Je remercie chaleureusement les membres du jury, Cécile Fabre et Alexandre Allauzen d'avoir examiné mon travail avec attention et de s'être déplacés. Leurs remarques m'ont été utiles. Je tiens à exprimer ma chaleureuse gratitude tout particulièrement à Cécile qui a apporté un suivi continu à mes recherches et une amélioration considérable au présent manuscrit grâce à ses remarques. Je suis heureux d'avoir pu bénéficier de points de vue aussi complémentaires sur mon travail.

Je remercie également, de la manière la plus vive, les collaborateurs du projet CRISTAL, dont les travaux ont contribué à faire progresser ce sujet délicat. Parmi eux, un remerciement tout spécial à Emmanuel Planas sans qui les expérimentations de ce travail n'auraient été réalisées. Je le remercie pour son suivi tout au long de cette thèse, son aide et ses réponses avec bonne humeur à mes questions quant au travail du traducteur.

Je remercie l'équipe TALN, spécialement Florian Boudin d'être régulièrement passé dans le bureau pour mettre de l'ambiance avec ses conseils [1] et autres plaisanteries coquines, et d'avoir été un redoutable partenaire sportif. J'en profite pour remercier chaleureusement mes collègues de bureau qui se sont succédés durant ces quatre années : Adeline, Amir, Grégoire, Hugo, Mohamed et Soufian. Je leurs souhaite une bonne continuation.

Je remercie tous ceux qui ont participé de près ou de loin à l'aboutissement de cet effort. Merci à tous les gens de l'ancien LINA ; administratifs, permanents et doctorants. Merci plus particulièrement à Amine, Evgeny, Walid, Hanen, Georges [2]... pour leur amitié. Une pensée à mes amis Anis, Amir et Soumaya qui sont également passés par cette étape.

J'adresse mes plus grands remerciements à ma famille ; mes parents, mon frère, ma sœur, Kaouther, mes nièces Joumana et Loujayn, ainsi qu'à Ibrahim, qui m'ont toujours soutenu dans mes choix, qui, je le sais, sont fiers de moi, et qui ont toujours été compréhensifs quant à mon manque de disponibilité.

1. Du type `#trop_de_boulot_tue_le_boulot`.
2. La liste est encore longue.

Pour finir, mes derniers remerciements vont à Anaïs [3]. Merci infiniment d'être toujours à mes cotés, d'avoir été aux petits soins quel que soit mon état de stress. Merci d'avoir compris que je ne pouvais pas anticiper mes vacances et être entièrement disponible, mais surtout d'avoir accompagné mes moments de bonheur. Je te souhaite bon courage pour la suite de ta thèse et tout le meilleur dans ta carrière.

3. Par qui je remercie également toute la famille Barateau.

# Table des matières

# Liste des figures

# Liste des tables

# Introduction 1

## 1.1 Contexte général

Depuis une vingtaine d'années, du fait de l'internationalisation, les entreprises se sont de plus en plus vues confrontées à la problématique du multilinguisme. Elles se sont trouvées face à de nouvelles contraintes d'organisation du travail, au titre desquelles nous citons en particulier l'utilisation d'outils communs à plusieurs implantations (logiciels et systèmes de gestion, équipes de travail multi-sites...), l'échange de l'information entre ces implantations, ou encore les relations entre sièges sociaux et filiales. Ainsi, les problèmes de langues sont liés au bon fonctionnement interne des entreprises. Il serait alors primordial pour chacune d'entre elles d'échanger avec ses partenaires, clients et employés dans leur langue maternelle, et de donner à ses employés la possibilité de communiquer aisément dans une langue étrangère, en particulier dans leur domaine d'expertise. Bien qu'il soit une richesse socio-culturelle indéniable, le multilinguisme, pose également de nombreux défis aux entreprises, notamment au niveau de leur économie. Le rapport ELAN (Hagen et al. 2006) estimait que le manque de compétences linguistiques avait fait perdre, sur une période de trois ans, une moyenne de 325 000 euros par PME européenne.

La maîtrise de la terminologie est un facteur considérable influençant l'organisation et l'économie de l'entreprise. Elle permet la circulation d'informations cohérentes tout en réduisant leur temps de production, et garantit un accès efficace aux contenus textuels et aux connaissances de l'entreprise. La nécessité de maîtriser la terminologie au sein d'un même groupe implanté dans différentes régions du monde, mais aussi la communication à l'international, et le souci de réaliser des économies conduisent le plus souvent les entreprises à privilégier l'usage de certaines technologies de gestion terminologique et d'aide à la traduction. En particulier, lorsqu'elle est orientée vers des utilisateurs non-experts, la mise à disposition d'outils d'aide à la traduction réduit l'insécurité linguistique qui est un facteur de stress au travail non négligeable. Ces technologies facilitent également les processus de localisation des contenus, et assurent à chacun un accès à l'information dans sa langue natale. Enfin, la maîtrise de la terminologie garantit l'emploi du mot juste, ce qui est crucial dans le cas de la rédaction de textes législatifs, médicaux ou encore de guides d'utilisation.

Dans cette perspective, de nombreuses recherches ont été consacrées

à l'accélération et l'amélioration de la traduction humaine via des logiciels de traitement automatique de l'écrit. Ces outils s'appuient principalement sur des ressources linguistiques propres à l'entreprise (moteurs de recherche, systèmes de traduction automatique, aide à la rédaction, de gestion de contenu, fouille de textes). Cependant, il est très rare de trouver des ressources linguistiques adaptées au domaine technique de l'entreprise. Une des raisons en est la « patte terminologique » de l'entreprise qui conserve sa terminologie propre. On parle ainsi de « *boite à gant* » chez Renault, mais de « *vide poche* » chez Citroën. Qui plus est, les connaissances techniques, tout comme les dénominations utilisées pour désigner ces connaissances, évoluent très rapidement ce qui rend caduque et financièrement inabordable toute tentative de création manuelle de telles ressources.

Dans ce cadre de recherche, s'inscrit cette thèse financée par le projet ANR-CRISTAL (ANR-12CORD-020). Il a été lancé en vue de développer un outil innovant capable d'extraire automatiquement à partir des textes de l'entreprise un véritable dictionnaire qui documente, traduit et facilite l'accès à toutes les expressions propres à l'entreprise et aux termes techniques liés à son domaine d'activité. Ce dictionnaire est ensuite valorisé à travers une plateforme dédiée à la traduction et à la gestion terminologique. Il servira d'outil de traduction assistée par ordinateur (TAO) et de gestion terminologique répondant aux besoins des entreprises et de ceux des organismes publics en terme de gestion de l'écrit monolingue et multilingue.

## 1.2 Motivations et objectifs

Le marché des technologies de la traduction intègre celui, plus confidentiel, des outils de gestion et de création de ressources linguistiques monolingues ou multilingues. Il s'est développé sur la base des différentes applications de traitement automatique de la langue (TAL) et comprend principalement les systèmes de gestion terminologique, les mémoires de traduction et les logiciels de traduction automatique. Beaucoup de solutions existantes sont en fait des solutions de niches, très souvent intégrées à des services plus génériques. Par conséquent, plusieurs outils technologiques touchant la sphère de la traduction se sont multipliés dans le poste personnel pour venir en aide aux traducteurs dans leur métier. Parmi les principaux acteurs du marché il existe à l'heure actuelle : Trados, Multitrans, Across, Déjà vu, Wordfast, MémoQ pour les **mémoires de traductions** ; Systran, IBM, Language Weaver, Reverso, Google pour la **traduction automatique** et Interverbum, Trados, Multicorpora, Terminotix pour la **gestion terminologique**.

La pénurie des compétences en traduction, favorisée par l'essor économique des pays émérgents, diversifie et augmente la demande en compétences linguistiques. Des études telles que Hagen et al. (2006) ont fait prendre conscience de l'impact de la pénurie en compétences linguistiques sur le développement économique des entreprises européennes, en particulier la pénurie des compétences en traduction avec les pays émérgents.

Si les traducteurs et autres professionnels de la langue sont les utilisateurs historiques des logiciels de TAO et de gestion terminologique, il se dessine aujourd'hui un nouveau pan d'utilisateurs constitué des collaborateurs ordinaires de l'entreprise, qui n'ont pas de formation spécifique en terminologie ou en langues étrangères. Cependant, les outils existants ne correspondent pas encore tout à fait aux attentes de ces nouveaux utilisateurs. Deux caractéristiques sont avancées. D'une part, ce ne sont pas des utilisateurs experts : ils ne maîtrisent pas le processus de traduction et n'ont pas pleinement conscience des difficultés de la tâche. D'autre part, ils n'acceptent pas que la recherche de termes ou de traductions soit laborieuse : par exemple, qu'ils doivent consulter plusieurs contextes avant d'obtenir une information précise ou intéressante.

Malgré la place importante qu'ils occupent dans l'aide à la traduction, les outils technologiques utilisés par les traducteurs ne sont pas encore capables de fournir toutes les informations contextuelles dont cet utilisateur aurait besoin. En effet, la principale difficulté dans l'exercice de traduction semble être liée à l'accès à cette information contextuelle qui est partiellement présente dans les outils d'aide à la traduction (Varantola 2006, Bowker 2011). Ces derniers, notamment les banques terminologiques, proposent peu d'informations et se focalisent le plus souvent sur les définitions et certaines relations pragmatiques comme la synonymie. Même s'ils commencent à être incorporés dans ces outils, les contextes proposés sous formes d'exemples illustratifs restent peu variés (Bowker 2011, p. 214) par rapport aux attentes du traducteur. En effet, lors d'une traduction très spécialisée (dans un sous-domaine par exemple), cette information doit être habituellement complétée par une recherche de contextes : localement dans le document à traduire, ou parcimonieusement dans des corpus complémentaires. Dans ces deux cas, les corpus sont exploités avec des outils plus ou moins adaptés (comme les concordanciers), ce qui constitue une difficulté supplémentaire au travail du traducteur.

Si l'accès aux expressions et aux termes techniques ainsi qu'à leurs traductions s'avère indispensable à tout processus de communication, leur potentiel doit être décuplé au moyen d'une « contextualisation » de ces termes et expressions. En effet, des travaux tels que celui de Durieux (1988) affirment qu'outre l'accès à un terme ou à sa traduction, encore faut-il être capable d'en appréhender le sens exact et de l'employer correctement. La contextualisation se manifeste ainsi à deux niveaux :

1. **conceptuel :** l'utilisateur doit avoir accès aux relations sémantiques (synonymie, antonymie...) ou conceptuelles (grâce à des structures telles que *est-un*, *est-une-partie-de*...) entre les termes afin de mieux en saisir le sens. Par exemple, il doit savoir que le terme *stigmatisme* évoque une propriété d'un *système optique*. Il faut donc pouvoir extraire automatiquement des *contextes riche en connaissances conceptuelles*.
2. **textuel :** l'utilisateur doit avoir accès à des informations concernant l'usage des termes. Par exemple, il doit savoir que le terme *laser* s'emploie avec le verbe *focaliser* lorsque l'on veut évoquer l'action

de concentrer les rayons du laser. Pour cela, il faut pouvoir extraire automatiquement des *contextes riches en connaissances linguistiques*.

Ces deux types de contextualisation ont pour but d'alléger considérablement le travail des terminologues et lexicographes spécialisés dont le métier est de renseigner, valider et maintenir les ressources linguistiques de l'entreprise, notamment sa terminologie. En améliorant l'ergonomie de travail des traducteurs et terminologues, ces contextes – dits « Contextes Riches en Connaissances » (CRC) (Meyer 2001) – répondent parfaitement à leurs besoins. En effet, ils mettent en relief l'information linguistique ou conceptuelle contenue dans les textes ce qui a pour effet d'accélérer le temps d'accès à l'information pertinente, et attirer l'attention de l'utilisateur sur des phénomènes linguistiques qu'il ne soupçonne pas. L'acquisition de cette technologie entre dans la stratégie globale du projet CRISTAL.

La technologie d'extraction automatique des CRC dans le cadre de CRISTAL nécessite la conception d'algorithmes d'extraction de connaissances capables d'analyser automatiquement et de façon robuste le contenu linguistique et sémantique de textes bruts et non-structurés et ce, qu'elle qu'en soit la langue et le domaine abordé. Les nouveaux dictionnaires proposés présenteront, pour chaque terme et ses traductions éventuelles, une fiche terminologique listant les CRC et explicitant les connaissances qu'ils contiennent. D'un point de vue scientifique, les retombées attendues sont la mise en œuvre d'une nouvelle génération d'outils d'aide à la traduction et à la gestion terminologique.

Du fait des informations précises et des exemples authentiques qu'ils contiennent, les corpus comparables semblent être considérés en pratique comme une matière première linguistique et terminologique appréciable également par les traducteurs (Williams 2008). Les travaux de cette thèse s'articulent autour de l'exploitation des corpus comparables spécialisés dans le but d'illustrer la terminologie par des CRC. Ces derniers auront pour rôle de faciliter la compréhension et l'usage de cette terminologie, en particulier dans la perspective d'aider à sa traduction : en **compréhension** et en **révision**.

La démarche que nous suivons consiste à étudier, dans un premier temps, les CRC du point de vue linguistique ainsi que leur apport par rapport à la compréhension en TAO : la possibilité d'appréhender le sens et l'usage de la terminologie en se basant sur les connaissances fournies par les CRC en langue source. Nous souhaitons associer pour chaque terme des CRC en langue source. Ces CRC seront tout d'abord évalués dans un cadre linguistique afin de vérifier leur validité, et ensuite par rapport à leur intérêt dans un exercice d'aide à la compréhension.

Dans un second temps, nous étudions l'apport des CRC dans le cadre de la révision en TAO. Nous nous intéressons ainsi à identifier des CRC bilingues équivalents qui illustrent une traduction attestée, c'est-à-dire un couple (terme et sa traduction proposée par le traducteur). Ces CRC bilingues auront pour but de confirmer ou infirmer cette traduction initialement attestée.

## 1.3 Plan de la thèse

La présente thèse se décompose en trois parties. Dans la première, le **chapitre 1** présente le contexte applicatif de nos recherches qui est la notion de « contexte » et son utilité pour les traducteurs, notamment en domaine de spécialité. Ensuite, le **chapitre 2** définit d'un point de vue linguistique les contextes riches en connaissances ainsi que les concepts liés à cette notion. Le **chapitre 3** contient l'état de l'art traitant de l'extraction des contextes qui peuvent manifester une certaine richesse en connaissances, principalement dans un cadre terminologique ou lexicographique.

La deuxième partie est composée de deux chapitres. Le premier (**chapitre 4**) rappelle tout d'abord la compréhension en traduction, puis motive et décrit notre contribution portant sur l'identification des CRC en vue d'aider le traducteur dans l'étape de compréhension. Ces CRC sont censés aider au mieux le traducteur à appréhender le sens et l'usage d'un terme donné dans son texte d'origine, facilitant ainsi sa traduction. Le second (**chapitre 5**) porte sur les ressources utilisées et les évaluations qui ont permis de valider nos hypothèses.

La troisième partie contient également deux chapitres. Dans le **chapitre 6** nous décrivons le cadre de la révision en traduction et motivons les hypothèses qui ont conduit à l'adoption de notre méthodologie. Celle-ci permet de construire un prototype de concordancier bilingue dédié à la validation d'une traduction produite en amont. Le **chapitre 7** expose les expérimentations réalisées ainsi que les résultats que nous avons obtenus.

Pour terminer, nous faisons le bilan de nos travaux et abordons nos perspectives de recherches, dans le dernier chapitre.

# Première partie

# Terminologie et état de l'art

# Contextes et traduction assistée par ordinateur

# 2

Sommaire

## 2.1 Introduction

Depuis l'utilisation massive du poste de travail informatique, la traduction assistée par ordinateur s'est développée sous forme d'un panel d'outils destinés à assister le traducteur dans son travail plutôt qu'à le remplacer. Différents outils se sont succédés dont les mémoires de traduction, les systèmes de traduction automatique, les logiciels d'extraction et de gestion terminologique, ainsi que les concordanciers multilingues. Les logiciels de mémoire de traduction s'appuient sur des documents initialement traduits qui constituent la base de ce qui sera plus tard connu sous le nom de corpus parallèles (Véronis 2000). L'objectif de ces logiciels est d'assister les traducteurs lors du processus de traduction en trouvant pour eux des passages préalablement traduits. Ils permettent également de recycler les traductions passées : lorsqu'un traducteur doit traduire une nouvelle phrase, le logiciel parcourt la mémoire à la recherche de phrases similaires préalablement traduites, et le cas échéant, propose la traduction passée comme modèle de traduction (Bowker et Barlow 2004).

Dans les années 90, les apports de la traduction assistée par ordinateur, de la traduction automatique et de ceux de la terminologie computationnelle (Bourigault 1994, Daille 1994, Jacquemin 1995) se sont entrelacés en donnant naissance aux algorithmes d'alignement de termes à partir de corpus parallèles (Daille et al. 1994, Melamed 1999, Gaussier et al. 2000). Les sorties de ces algorithmes sont des listes terminologiques bilingues, particulièrement utiles dans le cas de la traduction en domaine de spécialité. Ces dernières années, la qualité de la traduction automatique s'est améliorée, c'est pourquoi les traducteurs automatiques semblent être largement utilisés en traduction professionnelle. Aujourd'hui, ces outils donnent des résultats exploitables dans les domaines de spécialité dans lesquels le vocabulaire et les structures sont assez répétitifs (Delpech 2013). Il se peut parfois que le traducteur ait une intuition de traduction qui n'apparaît pas parmi les sorties du traducteur automatique qu'il utilise. Dans ce cas, les logiciels de gestion terminologique, qui ont évolué vers les banques terminologiques multilingues telles que TERMIUM ou UNTERM, lui permettent de chercher cette traduction potentielle dans le corpus souhaité. Les concordanciers multilingues constituent également une aide précieuse au traducteur en lui donnant l'accès aux contextes d'un mot (ou d'un terme) et mettent en regard la traduction de ces contextes en langue cible.

Le poste de travail du traducteur est à présent un environnement qui est devenu fortement informatisé et intégrant de nombreux logiciels comme Trados, Wordfast, DéjaVu pour en citer quelques uns. D'autres logiciels à destination du grand public sont aussi largement exploités par les professionnels. Les plus populaires étant Linguee [1] pour les concordanciers bilingues, et celui de Google : Google Translate [2] pour les systèmes de traduction automatique. Franz Och, chercheur chez Google Translate, es-

1. http://www.linguee.fr/
2. www.translate.google.com

time que l'équivalent du contenu d'un million de livres est traduit chaque jour en ligne sur ce site. Même si l'utilisation de ces services en ligne n'est pas recommandée dans un cadre professionnel, ces systèmes viennent répondre à un besoin toujours croissant de compréhension d'informations en langues étrangères. Cependant, les limites de ces outils sont rapidement atteintes dès qu'il s'agit d'une traduction très spécialisée. Par exemple, pour traduire de l'anglais en français le terme *blob* provenant du domaine de la vulcanologie, GoogleTranslate propose *goutte*, *tache* et *pâté* (figure 2.1) avec deux définitions suivies de deux exemples que l'on peut juger « peu utiles » dans ce cas. Concernant Linguee, il propose également *pâté*, *goutte* et *blob*, mais avec plus de contextes qui ne contiennent pas la traduction correcte : *projection* (figure 2.2).


FIGURE 2.1 – *Résultat de traduction de blob avec GoogleTranslate*

Si l'on envisage le processus de traduction lui-même comme « une enquête non linéaire » tel que décrit par Rochard (1999), le traducteur n'est pas un simple « passeur » d'une langue à l'autre, il doit pêle-mêle mettre en évidence les éléments logiques du texte original, comprendre la signification de l'énoncé en texte source, le déverbaliser, se référer à des informations textuelles (du monde réel), choisir une formulation adéquate en langue cible ayant le même niveau de performance fonctionnelle que celle du texte source. L'objectif de cette thèse se focalise sur ce type d'information « contextuelle » qui pourrait alléger et améliorer la qualité du travail du traducteur, particulièrement en domaine de spécialité.

## 2.2 Notion de contexte pour les traducteurs

De nombreux travaux se sont penchés sur l'étude du comportement des traducteurs afin d'améliorer l'enseignement de la traductologie ou le fonctionnement des outils d'aide à la traduction (Varantola 1998, Künzli 2001, Bowker 2008, Désilets et al. 2009). Dans ce type d'études, l'un des objectifs est d'identifier les extraits de ressources qui focalisent l'attention des traducteurs. Il s'agit de passages de textes particulièrement utiles au sens


FIGURE 2.2 – *Résultat de traduction de blob avec Linguee*

où ils jouent un rôle positif dans l'exercice de traduction. Par conséquent, lorsque l'on souhaite fournir aux traducteurs des outils qui répondent mieux à leurs besoins, on devrait en premier lieu se demander ce qui rend un contexte pertinent pour eux. En d'autres termes, ce qui pourrait être un « bon contexte » pour les traducteurs. Selon Rogers et Ahmad (1998, p. 195), l'un des premiers besoins du traducteur est l'information sensible au contexte (en anglais *context-sensitive information*). Nous nous intéressons alors à ce que cette notion d'information sensible au contexte pourrait englober et aux sources d'informations qui ont du succès auprès des traducteurs.

Bien qu'elle soit largement reconnue comme étant essentielle pour le traducteur, la notion de contexte, en traduction, reste difficile à définir (Baker 2006), et manque de définition pratique qui pourrait être appliquée dans le travail quotidien des traducteurs professionnels (Melby et al. 2010, p. 1). Melby et al. (2010) ont finement détaillé comment des spécialistes dans de nombreux domaines (philosophie, psychologie, pragmatique et linguistique fonctionnelle) ont traité la notion de contexte pour aboutir à diverses propositions de définitions. Les trois facettes du contexte telles que définies par Halliday (1999), à savoir le contexte de la situation, le contexte de la culture et le co-texte, sont particulièrement pertinentes dans la traduction. Contrairement au contexte de la situation et celui de la culture, qui sont en dehors du langage lui-même [3], le co-texte se rapporte spécifiquement à la langue utilisée. Il peut largement être défini comme le discours environnant (ou aux alentours) d'un énoncé donné. Par conséquent, notre définition de « contexte » dans le présent travail sera limitée

3. Ces contextes désignent l'ensemble des circonstances dans lesquelles se produit un acte d'énonciation tel que la situation culturelle et psychologique, expériences...

à ce que Halliday appelle co-texte, et se fondera sur la définition donnée par Fuchs [4] :

> « *On appelle 'contexte' l'entourage linguistique d'un élément (unité phonique, mot ou séquence de mots) au sein de l'énoncé où il apparaît, c'est-à-dire la série des unités qui le précèdent et qui le suivent : ainsi, dans l'énoncé 'Marie est jolie comme un cœur', l'élément 'comme' a pour contexte immédiat 'jolie... un cœur' et pour contexte plus large l'environnement 'Marie est jolie... un cœur'. Par extension, on parle également du contexte d'un énoncé au sein d'un discours pour désigner le ou les énoncés qui précèdent et suivent immédiatement l'énoncé considéré.* »

## 2.3 Contextes utiles aux traducteurs

Roberts et Bossé-Andrieu (2006, p. 203) soulignent que les traducteurs font face à des problèmes qui peuvent être principalement liés aux textes sources, notamment au niveau de la compréhension du texte à traduire ; ou aux textes cibles pendant la transition vers la langue cible. L'auteur a regroupé ces problèmes en trois principales catégories : **encyclopédique**, **linguistique** et **textuelle**. Les problèmes encyclopédiques couvrent les problèmes liés aux sujets généraux ainsi que des problèmes plus spécifiques portant sur des noms propres n'étant pas familiarisés avec le « sujet » du texte. Les problèmes linguistiques sont associés à des mots ou des phrases spécifiques. Enfin, les problèmes textuels concernent les types de textes tels que l'organisation interne et la production d'un type particulier de texte.

Bowker (2011; 2012) établit une liste des éléments portant sur les informations contextuelles qui peuvent se révéler « utiles » pour le traducteur dans la résolution des problèmes liés aux textes sources et cibles. Josselin-Leray et al. (2014) ont résumé ces éléments comme suit :

a) des informations sur l'usage incluant, bien entendu, les collocations, en particulier les mots de langue générale qui co-occurrent avec des termes (Roberts 1994, p. 56) ;
b) des informations sur la fréquence d'utilisation d'un mot ou d'un terme particulier ;
c) des informations sur les relations lexicales et conceptuelles comme la synonymie, méronymie, hyperonymie, etc. (Marshman et al. 2012, Rogers et Ahmad 1998) ;
d) des informations pragmatiques sur le style et le genre (Varantola 1998) ;
e) des informations sur les usages à éviter.

Il nous semble alors que ces éléments, qui ne sont liés ni à la situation ni à la culture, peuvent être classés dans les catégories suivantes : conceptuelles (c) et linguistiques (a, b, d et e). Afin de résoudre les problèmes et

4. http://www.universalis.fr/encyclopedie/contexte-linguistique/

de prendre des décisions, les traducteurs ont besoin d'aide, obtenue généralement en sollicitant d'autres experts ou en consultant des ressources plus conventionnelles telles que les dictionnaires et les banques terminologiques (Rogers et Ahmad 1998, p. 198).

## 2.4 Accès aux ressources en traduction

Nous présentons dans cette section les ressources conventionnelles utilisées en traduction et nous discutons l'utilité des connaissances qu'elles procurent par rapport au besoin du traducteur.

### 2.4.1 Ressources en traduction

En complément des dictionnaires monolingues et bilingues ainsi que les banques terminologiques, les traducteurs s'appuient partiellement sur les informations provenant de corpus. Nous détaillons ci-après ces ressources.

— **Les dictionnaires :** c'est le plus souvent à travers les exemples que les dictionnaires fournissent des informations contextuelles. L'étude empirique réalisée par Josselin-Leray (2005) sur les termes en dictionnaires généraux (monolingues et bilingues) montre que plus de 80,3 % des utilisateurs ont recours aux dictionnaires afin de trouver des informations concernant l'usage des termes dans des phrases, et que les dictionnaires bilingues ont plus de succès. La majorité des personnes interrogées ayant choisi cette réponse, sont des « professionnels de la langue » comprenant également des traducteurs.

Roberts (1994, p. 56) a mené une enquête auprès des utilisateurs potentiels du Dictionnaire Canadien Bilingue afin d'identifier leurs besoins. En conclusion, entre un tiers et la moitié des membres de chaque groupe d'utilisateurs ont apprécié, à différents degrés, le nombre d'exemples présentés dans leurs dictionnaires habituels. Cependant entre un quart et la moitié de chaque groupe a estimé qu'une amélioration était nécessaire. L'étude réalisée par Josselin-Leray (2005) a abouti à la même conclusion : bien que les utilisateurs aient été globalement satisfaits par les exemples fournis par leurs dictionnaires (entre 41,3 % et 67,5 % des utilisateurs ont déclaré qu'ils étaient satisfaits), le niveau de satisfaction était plus faible pour les dictionnaires bilingues, et encore plus faible pour le groupe d'utilisateurs « professionnels de la langue ».

— **Les banques terminologiques :** elles fournissent des informations contextuelles inscrites dans le champ « contexte » de la fiche terminologique. L'importance associée par les traducteurs à l'information contextuelle, notamment en ressources terminologiques, a été confirmée par les résultats de l'enquête menée par Duran-Muñoz (2010) : les exemples étaient considérés comme des « données essentielles » par les personnes interrogées, parmi les « données désirables ».

Dans le sondage réalisé par Duran-Muñoz (2010), les traducteurs ont eu l'occasion de donner leur avis au sujet de leurs besoins en traduction. Le second [5] argument le plus fréquent a été d'« inclure des informations plus pragmatiques sur l'usage et sur les traductions difficiles », et le cinquième de « fournir des exemples tirés de textes réels ». Dans la conclusion de ses travaux, Munoz (2012, p. 82) confirme clairement que la plupart des ressources terminologiques actuellement disponibles (en particulier sous forme électronique) ne satisfont pas les exigences des traducteurs.

— **Les corpus :** bien que les corpus soient intrinsèquement construits d'informations contextuelles, ce sont des ressources qui semblent encore rarement utilisées par les traducteurs, comme le montre les résultats de l'étude Duran-Muñoz (2010) : seulement 5,09 % des participants ont cité les corpus (parallèles en particulier) comme étant une ressource terminologique qu'ils utilisent plus de 4 fois lors de la traduction. Cependant, 41,8 % des personnes interrogées à la Mellange Survey [6], réalisée en 2005-2006 avec des traducteurs stagiaires et professionnels, prétendent utiliser les corpus dans leur pratique de traduction, notamment les corpus en langue cible.

Bien qu'il semble y avoir un large éventail de ressources fournissant de l'information contextuelle, ces ressources ne répondent pas nécessairement aux besoins des traducteurs.

### 2.4.2 Terme hors contexte *vs.* terme avec contexte

Varantola (1998, p. 180) souligne que les dictionnaires, qui sont le principal outil utilisé par les traducteurs, ont un rapport très problématique avec la notion de contexte. Tandis que les dictionnaires cherchent à faire des descriptions de mots généralistes détachés de tout contexte spécifique, les traducteurs cherchent à résoudre des problèmes très dépendants de domaines de spécialité. Le problème de la longueur des unités de traduction se pose également : le dictionnaire présente les données sous forme d'unités lexicales souvent courtes, alors que les traducteurs privilégient des unités textuelles plus longues qu'un seul mot.

Varantola (1998) a étudié de manière approfondie le dilemme *context-free vs. context-bound* rencontré par les traducteurs : les dictionnaires et les banques terminologiques ne fournissent que des exemples hors contextes (*context-free*) perçus comme prototypiques et fréquents ; alors que ce dont les traducteurs ont besoin pour trouver la meilleure traduction en langue cible est typiquement liée au contexte (*context-bound*). En outre, les exemples (*i.e* contextes) fournis ne sont pas assez variés, en particulier dans les banques terminologiques. Cela a également été mentionné par Bowker (2011, p. 214) qui explique que l'information trouvée dans les fiches terminologiques est habituellement limitée à des définitions et des termes présentés hors contexte ou dans un seul contexte dans le meilleur

5. Ici, nous citons seulement les arguments en relation avec la suite de notre travail.
6. `http://mellange.eila.jussieu.fr/Mellange-Results-1.pdf`

des cas. L'auteur met en évidence une situation paradoxale dans laquelle les progrès de la recherche sur la terminologie n'ont pas été intégrés dans les outils que les traducteurs utilisent le plus souvent : les banques terminologiques.

Varantola (2006, p. 217) affirme également que le dilemme *context-free vs. context-bound* doit être résolu du fait de la disponibilité de grands corpus, dont le rôle est désormais central dans la compilation des dictionnaires :

> « *Context-free definitions of concepts within a particular domain [which] were for a long time the theoretical ideal in terminological theory [...] are now replaced by less rigid, contextually relevant definitions.* »[7]

Certains dictionnaires donnent maintenant accès à d'autres exemples sous forme de concordances à partir des données de corpus qui se cachent derrière le dictionnaire. Cependant, Varantola (2006, p. 223) reproche aux corpus le fait qu'ils soient des outils d'intelligence superficielle (*tools of shallow intelligence*) lorsqu'ils sont à l'état brut et non étiquetés en parties du discours. Dans ce cas, l'utilisateur doit à la fois gérer la manipulation, la dissection et l'interprétation des résultats de ses recherches. En d'autres termes, la compilation et l'analyse des corpus peuvent être une tâche trop fastidieuse pour les traducteurs qui travaillent souvent sous de fortes contraintes, notamment temporelles.

## 2.5 Outil « idéal » en traduction

En 1996, Atkins (1996, p. 526) a proposé que le dictionnaire de l'avenir devrait offrir à ses utilisateurs la possibilité de prendre leurs propres décisions concernant les unités équivalentes dans les deux langues :

> « *[They] should be able to consult as many examples as they need of words used in their various senses, each in a variety of contexts with a variety of collocate partners* »[8](Atkins 1996, p. 526)

Plus récemment, Bowker (2011, p. 215) a suggéré d'améliorer les ressources terminologiques comme suit :

> *"It would be more helpful for translators to have access not simply to term records that provide a signle 'best' term with a solitary context, but rather to information that would allow them to see all*

7. « *Les définitions hors-contextes de concepts pour un domaine particulier [qui] ont été pendant longtemps l'idéal théorie en terminologie théorique [...] sont maintenant remplacés par des définitions moins rigides, contextuellement pertinentes.* »

8. « *[Ils] devraient être en mesure de consulter autant d'exemples que nécessaire, car ils ont besoin de termes utilisés dans leurs différents sens, chacun dans une variété de contextes avec une variété de collocatifs* ».

> *possible terms in a range of contexts and thus find the solution that works best in the target text at hand"* [9] Bowker (2011, p. 215)

Ici, Bowker (2011) insiste sur le fait que la recherche dans un grand éventail de contextes ne doit pas être considérée comme une perte de temps. Cela a été rendu plus facile grâce à des outils d'analyse de corpus qui présentent l'information dans un format facile à lire. Bowker (2012, p. 391) va même plus loin en suggérant de donner aux traducteurs l'accès à l'ensemble des informations qu'utilisent habituellement les lexicographes lors de l'élaboration des entrées de dictionnaires :

« *In order to arrive to that entry, lexicographers have gone through a number of intermediary steps, where they learn about the various characteristics of the words and concepts being described, such as their grammatical and collocational behaviours, the different relationships that hold between words and their underlying concepts, and the characteristics that are necessary and sufficient for distinguishing one concept in an intensional definition.* » [10]

Les idées que portent les lexicographes et terminologues sur la notion de « bons exemples » fournissent des informations précieuses sur ce que pourrait être un bon contexte pour les traducteurs.

## 2.6 Conclusion

De nombreuses études ont souligné l'importance du coté illustratif des dictionnaires comme étant un moyen de fournir des contextes typiques concernant le sens et l'usage des mots (Atkins et Rundell 2008). Dans les dictionnaires monolingues et bilingues, des exemples ont pour but d'aider l'utilisateur à la fois dans la compréhension de la langue source et la production de traductions en langue cible. Ces dictionnaires auront ainsi différentes utilités : ils peuvent fournir des informations syntagmatiques sur les termes telles que leurs voisinage d'apparition et leurs collocations, ainsi que des informations paradigmatiques illustrant des relations sémantiques entre les termes (synonymes, hyperonymes, etc.). L'ensemble de ces indications construit des connaissances pragmatiques et stylistiques sur les usages spécifiques. Concrètement, ces contextes pourront être accessibles comme un complément aux définitions, avec une dimension épilinguistique. Une problématique majeure tient au fait que ces exemples authentiques, qui répondent à certaines exigences des traducteurs, n'ont pas été clairement définis dans l'état de l'art et sont ainsi difficiles à extraire en corpus de spécialité.

9. « *Il serait plus utile pour les traducteurs d'avoir accès non seulement aux fiches des termes qui fournissent un seul 'meilleur' terme avec un contexte solitaire, qu'aux informations qui leurs permettraient de voir tous les termes possibles dans une série de contextes et ainsi trouver la solution qui fonctionne le mieux dans le texte cible en question.* »

10. « *Afin de produire cette entrée, les lexicographes sont passés par un certain nombre d'étapes intermédiaires, dans lesquelles ils apprennent les différentes caractéristiques des mots et des concepts étant décrits, tels que leurs comportements grammaticaux et collocationels, les différentes relations entre les mots et leurs concepts et les caractéristiques qui sont nécessaires et suffisantes pour distinguer un concept dans une définition intentionnelle.* »

# Indices et marqueurs de connaissances

3

Sommaire

## 3.1 Introduction

De nombreux travaux tels que ceux de Sager (1990) et Temmerman (2000) considèrent la la terminologie comme une approche interdisciplinaire à la croisée de différentes (sous)-disciplines : lexicologie, sémantique, linguistique cognitive, traduction, sociolinguistique, philosophie, etc. En traduction, l'identification de la terminologie et de son équivalent constitue une part importante du métier de traducteur, et a un impact majeur sur la qualité du document produit. Le client préfère utiliser la « patte terminologique » de son domaine de spécialité ou de sa compagnie. Pour une entreprise de meubles, par exemple, les termes *bureau* et *table* sont différents. Pourtant, un traducteur ignorant la terminologie de l'entreprise pourrait confondre ces deux termes en traduisant le terme anglais *desk*. Se croisent alors l'intérêt de la terminologie et celui de la traduction, et dans ce cas, le traducteur joue également un rôle de terminologue pour résoudre ce type de dilemme. Pour cela, une gestion efficace de la terminologie améliorera la qualité du document traduit et évitera des problèmes liés à la terminologie (*treminology-related task*) (Bowker 2003, p. 49). Aujourd'hui, plusieurs outils et ressources sont à la disposition des traducteurs, notamment les ressources termino-ontologiques.

Au début des années 2000, la notion de ressources termino-ontologiques (Bourigault et al. 2004) est apparue dans le sillage des travaux effectués sur les bases de connaissances terminologiques. Ces dernières ont été développées au début des années 1990 (Meyer et al. 1992, Condamines et Amsili 1993) en mettant l'accent sur le lien entre l'ingénierie des connaissances et la pratique terminologique. Avec ce changement de dénomination, la dimension informatique a été davantage mise en avant. En effet, la construction des réseaux terminologiques est passée de l'exploitation de l'expertise des terminologues, à des méthodes qui reposent sur l'exploitation de ressources (étant le plus souvent des textes perçus comme des réservoirs de connaissances), avant de solliciter les terminologues pour valider les connaissances identifiées. En traduction comme en terminologie, l'analyse du fonctionnement des termes dans les textes se base sur l'étude des contextes dans lesquels ils apparaissent. Pour cela, de nombreux spécialistes de la traduction ont souligné les mérites de la *traduction assistées par corpus*, appelée en anglais *corpus-aided translation*. Les corpus sont le complément idéal des ressources linguistiques conventionnelles qui sont très souvent incomplètes ou décontextualisées (Bennison et Bowker 2000). Ils répondent idéalement aux problèmes notamment soulevés par la traduction spécialisée, et peuvent contenir des exemples authentiques et des connaissances essentielles au travail de traduction. Afin d'acquérir ces éléments, les traducteurs se doivent de consulter les extraits de textes susceptibles de véhiculer les informations utiles via des outils et des ressources facilitant l'accès aux corpus utilisés.

Ce besoin a fait émerger la notion de « contextes riches en connaissances » introduite par Meyer (2001). Initialement définie dans un cadre terminologique, cette notion interroge les corpus textuels sur des relations

conceptuelles entre les termes en remettant en question les liens entre sens et connaissance, et entre forme linguistique et sens. Cette notion fait écho à d'autres types de contextes qui peuvent aussi être considérés comme « riches en connaissances », selon le sens que l'on donne à « connaissances » : nous parlons d'exemples lexicographiques. Ces derniers interrogent les textes plutôt sur des phénomènes linguistiques, en particulier sur le fonctionnement linguistique et les usages typiques souvent illustrés par les collocations.

Dans les prochaines sections nous présentons la terminologie qui sera principalement utilisée dans la présente thèse. Nous introduirons les contextes riches en connaissances, les marqueurs de relations et les patrons de connaissances, ainsi que la relation entre ces trois concepts. Nous aborderons également les collocations et les connaissances qu'elles peuvent procurer.

## 3.2 Contextes riches en connaissances

L'idée que certains contextes peuvent se révéler utiles pour appréhender le sens est loin d'être nouvelle. Aristote l'avait déjà abordée en affirmant l'existence de contextes définitoires. Aujourd'hui, l'évolution des ressources textuelles, désormais accessibles sous format numérique, a entraîné le besoin de décrire plus systématiquement ces contextes afin d'être en mesure de les récupérer automatiquement. Plus récemment, ce besoin s'est particulièrement concentré sur des domaines spécialisés pour construire des réseaux de termes (Williams 1998, Gillam et al. 2005).

Dans ce cadre, Meyer (2001) fut la première à proposer l'appellation contextes riches en connaissances (CRC) pour désigner les contextes qui permettent de repérer, grâce à des éléments lexico-syntaxiques, des relations (souvent lexicales ou lexico-syntaxiques) entre plusieurs termes du même domaine. Il s'agit de portions de textes qui contiennent *i*) **des termes** d'un domaine spécialisé et *ii*) **des marqueurs** explicitant des relations entre ces termes.

> "*By knowledge-rich context, we designate a context indicating at least one item of domain knowledge that could be useful for conceptual analysis. In other words, the context should indicate at least one conceptual characteristic, whether it be an attribute or a relation.*" [1]
> (Meyer 2001, p. 281)

Par exemple, la phrase *Les graisses dans le sang sont essentiellement le cholestérol et les triglycérides* est un contexte riche en connaissances pour le terme *graisse dans le sang*. Dans cette phrase, le marqueur *être_essentiellement_un* explicite une relation hiérarchique entre les termes *graisse dans le sang*, *cholestérol* et *triglycéride* issus du domaine médical.

1. « *Par contexte riche en connaissances, nous designons un contexte indiquant au moins un item du domaine des connaissances exploitable pour une analyse conceptuelle. Autrement dit, le contexte doit indiquer au moins une caractéristique conceptuelle, que ce soit un attribu ou une relation* »

Du fait des connaissances qu'ils illustrent, les CRC sont considérés très précieux et utiles pour l'acquisition de relations sémantiques entre les termes. L'exemple précédent contient une relation de définition (par dénotation) entre *graisse dans le sang*, *cholestérol* et *triglycéride*.

## 3.3 Marqueurs de relations

D'emblée, lorsque la notion de contexte riche en connaissances a été proposée, son lien avec les marqueurs de relations a été posé, au point que souvent, de manière beaucoup trop hâtive, les deux notions sont confondues. En effet, dans la littérature, les CRC tels que définis par Meyer (2001) sont le plus souvent identifiés grâce à des marqueurs de relations. Ceux-ci permettent de repérer et classer finement les relations terminologiques dans les corpus spécialisés. Il s'agit de mots, d'expressions ou de symboles révélant de façon récurrente une relation terminologique. Par exemple, la structure *telle que* est un marqueur de relation qui exprime une relation d'hyperonymie pouvant relier deux termes comme dans la phrase : *Une hormone telle que l'insuline...* dans laquelle *hormone* et *insuline* sont deux termes médicaux.

## 3.4 Patrons de connaissances

Pearson (1998) et Meyer (2001) ont montré l'intérêt des marqueurs linguistiques indiquant des relations sémantiques entre des termes pour exploiter les corpus spécialisés. Les marqueurs de relations sont modélisés et mis en œuvre grâce aux patrons de connaissances (PC). L'une des principales stratégies utilisées dans le but d'isoler les CRC, et ainsi d'écarter les contextes jugés moins utiles, consiste à utiliser les PC. Cette démarche permet aux terminologues de pointer particulièrement vers le sous-ensemble des phrases susceptibles de véhiculer les informations souhaitées (Barrière 2004). Un PC est une expression régulière formée de mots, de catégories grammaticales ou sémantiques et de symboles, visant à identifier des fragments de texte qui explicitent des marqueurs de relations. Par exemple, dans la phrase *X est un type de Y* (*X* et *Y* étant deux termes différents), la structure *est un type de* est un patron de connaissances modélisant le marqueur *être_DET*[2]*_type_DET*. Nous appelons PC définitoire, un PC s'articulant autour d'un marqueur de définition.

**Synthèse :** les contextes riches en connaissances sont des contextes contenant un patron de connaissances associé à un marqueur de relation. Reprenons l'exemple précédent :

- CRC : *les graisses dans le sang sont essentiellement le cholestérol et les triglycérides*.
- PC : *X sont essentiellement Y et Z* ; X, Y et Z sont trois termes.
- Marqueur de relation : *être_essentiellement*.

2. DET : déterminant.

## 3.5 CRC et définitions

Selon Meyer (2001), les CRC peuvent servir de point de départ à la construction de définitions de termes. De ce point de vue, ils se rapprochent fortement des « énoncés définitoires »[3] (Rebeyrolle 2000b). Ces énoncés peuvent être de différents types correspondant chacun à une structure linguistique pouvant être transcrite sous forme de marqueur de relation. Rebeyrolle et Tanguy (2000) qualifient ces énoncés, et plus généralement les définitions, d'objets particulièrement riches sur le plan sémantique.

### 3.5.1 Types de définitions

Bon nombre de travaux se sont intéressés aux définitions dans les textes spécialisés (Hermans 1989), (Flowerdew 1992b), (Flowerdew 1992a), (Pearson 1998), (Rebeyrolle et Tanguy 2000) et (Rebeyrolle 2000b). Tous ces travaux mentionnent l'apparition récurrente des marqueurs de relations dans les définitions. De manière générale, définir un terme consiste à en expliquer le sens dans un domaine précis. Selon Sager (1990), la définition est une description linguistique d'un concept. Elle doit respecter des règles assez strictes telles que l'absence de circularité[4]. Il existe trois principaux types de définitions.

— **Définition terminologique :** elle décrit un concept désigné par un terme, et le caractérise par rapport à d'autres concepts à l'intérieur d'un système structuré (appelé système conceptuel). Autrement dit, elle doit illustrer le plus clairement possible les relations entre les termes dans un domaine de spécialité bien précis. Ce type de définition décrit le concept représenté par une dénomination (Cabré 1998) en utilisant le patron *terme à définir_concept_référent* (Larivière 1996).
**ex.** *Clavier : périphérique d'entrée muni de touches alphanumériques et de touches de fonction, servant à saisir des données et à lancer des commandes.*

— **Définition lexicographique :** elle énumère et explique tous les sens (significations) du mot à définir. « *Elle se présente comme une périphrase synonymique du défini exhibant les éléments constituant le sens du mot défini. Ces éléments sont choisis par le lexicographe comme ceux qui sont susceptibles de faire comprendre le mot défini* » (Fradin et Marandin 1979). Cette définition se distingue de la définition terminologique par l'unité à définir et les procédés employés.
**ex.** *Clavier : ensemble des touches de certains instruments de musique (piano, orgue, accordéon, etc.), d'une machine à écrire, d'un terminal informatique, etc.*

— **Définition encyclopédique :** elle détaille un ensemble de connaissances en sciences humaines (linguistique, histoire...) qui concernent

3. Ce terme est utilisé par les auteurs pour désigner les contextes contenant un acte de définition dans un discours.

4. Définir un terme par un autre terme (ou concept) à définir.

le mot en question. Les définitions proposées par Wikipedia [5] sont le plus souvent encyclopédiques.

Si les définitions lexicographiques et encyclopédiques élargissent la signification que peut avoir un terme, la définition terminologique en détermine un seul sens. Nous nous intéresserons par la suite à la définition terminologique étant la plus proche de la notion de CRC, notamment en domaine de spécialité.

### 3.5.2 Typologies des définitions

De nombreux auteurs ont abordé les définitions terminologiques dans le texte spécialisé. Pascual et Péry-Woodley (1995), par exemple, ont examiné la structure et les divers composants linguistiques des définitions. La typologie regroupe trois formes différentes de définitions terminologiques, classées selon des critères bien définis. Ces trois typologies avaient comme objectif premier d'aider à la rédaction de définitions terminologiques, mais elles ont été également exploitées comme « indices » servant à repérer les définitions terminologiques et distinguer leurs composants.

Les travaux de Sager (1990), Nagao (1992) et Trimble (1985) présentent une littérature riche portant sur les typologies des définitions terminologiques. Trimble (1985) a présenté une méthode de rédaction de définitions destinée à des étudiants non natifs. Son travail illustre une variété d'expressions aidant à construire des définitions de termes. Nagao (1992) a quant à lui consacré son travail à détailler les notions des termes qui sont souvent abstraites, et ce dans des dictionnaires terminologiques.

Dans cette section, nous nous limitons à la typologie de Sager (1990) qui traite des définitions se manifestant dans les données terminologiques. Selon Sager (1990), les définitions terminologiques ont pour but d'être utilisées comme une « banque de données » par des traducteurs, spécialistes et amateurs. L'auteur a classé les définitions terminologiques en sept catégories présentées ci-après avec des exemples empruntés à l'auteur.

1. **Définition par analyse :** elle contient la classe du terme (genus) et la différence (differentia) qui le distingue des autres termes.
   **Ex.** *pneumonia = an inflammation (classe) of lung tissue (differentia).*
2. **Définition par synonyme :** elle illustre un ou plusieurs termes partageant la même signification. Elle peut également contenir un terme équivalent dans une autre langue.
   **Ex.** *software = logiciel.*
3. **Définition par paraphrase :** elle reformule le terme défini en décrivant son signifié avec des mots usuels.
   **Ex.** *whiteness = the state of being white.*
4. **Définition par synthèse :** elle donne les relations sémantiques entre les termes.
   **Ex.** *metatarsalgia = a painful neuralgic condition of the foot, felt in the ball of the foot and often spreading thence up the leg.*

5. www.wikipedia.com

5. **Définition par implication :** elle propose un exemple (contexte) contenant le terme.
   **Ex.** *dial = a clock or watch has a dial divided into segments for hours and minutes over which the hands move.*
6. **Définition par dénotation :** elle liste des exemples du terme.
   **Ex.** *dog = dogs are spaniels, poodles, Pekinese, Alsatians and similar animals.*
7. **Définition par démonstration :** elle illustre une image, un modèle ou un exemple concret du terme en question.

Les limites séparant les définitions des CRC n'ont pas été clairement décrites dans les travaux de l'état de l'art. En analysant les typologies des définitions terminologiques, nous constatons qu'elles correspondent, en pratique, à des formes strictes structurant les connaissances qui doivent être illustrées par les définitions. Ces typologies, qui peuvent également être perçues comme des patrons de connaissances, voire des CRC, aident à construire et formuler des définitions. Nous distinguons ainsi les CRC des définitions par le fait qu'ils sont des expressions plus ou moins bien structurées ne respectant pas de formes « canoniques », mais qui contiennent des connaissances aidant à construire des définitions. Une représentation de la typologie par analyse, par exemple, pourrait être modélisée sous la forme :

$$Terme_1 + PC\,hyperonymie + Hyperonyme + patron\_de\_distinction + Terme_2$$

## 3.6 Collocations et connaissances linguistiques

Du fait de leur utilité dans la construction de définitions et de réseaux terminologiques, les CRC tels que définis par Meyer (2001), visent uniquement les connaissances conceptuelles (c'est-à-dire les relations entre les termes). En revanche, les lexicographes manifestent un intérêt particulier à d'autres types de connaissances portant plutôt sur le fonctionnement linguistique des mots et leurs comportement dans leurs contextes d'apparition. Une des notions les plus appréciées en lexicographie est les collocations.

### 3.6.1 Collocations

Même si elles ne constituent pas un objet traditionnel de la terminologie, les collocations ont été progressivement prises en compte dans certains travaux à visée terminologique. Ces dernier induisent de nouvelles pratiques dans le recensement terminologique (Pavel 1993). Parmi les travaux les plus saillants, nous citons (Williams 1998) dans le domaine de la biologie végétale, ou (L'Homme 2008) dans celui de l'informatique. Williams (1998) s'intéresse aux co-occurrences significatives entre deux lexèmes, non seulement afin d'extraire des expression polylexicales mais aussi pour déterminer leur « rôle thématique, facteur de cohésion textuelle », ce qui l'amène à utiliser la notion de *réseaux de collocations*.

L'accès à des informations relatives à la combinatoire lexicale constitue une des priorités du traducteur lorsqu'il consulte des ressources terminologiques ou textuelles, voire lexicographiques. Ces informations correspondent, pour le traducteur, à une dimension importante de la connaissance qu'il vise à acquérir ou vérifier par la consultation de contextes, qu'il s'agisse de la langue générale ou d'usages spécialisés.

La maîtrise des collocations est une composante essentielle de la maîtrise de la langue ou d'un discours spécifique. Ceci explique l'importance accordée à cette notion que ce soit pour traduire un texte ou pour élaborer une ontologie. Au sens restreint, les collocations représentent des associations lexicales transparentes du point de vue de la compréhension mais qu'un locuteur non natif doit tout particulièrement apprendre à maîtriser. Ce qui est le cas des exemples : *prescrire une ordonnance*, *tenir debout*, *nuit blanche*.

Même si elle reçoit des définitions variables selon le contexte de recherche dans lequel elle est employée, la notion de collocation est habituellement définie selon deux traditions : TAListe ou linguistique. Sinclair et al. (1970) a proposé une définition TAListe, basée sur la co-occurrence de deux éléments :

> *"[...] the occurrence of two items in a context within a specified environment. Significant collocation is a regular collocation between two items, such that they co-occur more often than their respective frequencies and the length of the text in which they appear would predict."* [6] (Sinclair et al. 1970, p. 150)

Cette définition a fait émerger des propriétés abordées principalement dans les travaux de Hausmann (1989) et Mel'ćuk (1998). Ces propriétés permettent de donner une définition linguistique de la notion de collocation.

### 3.6.2 Propriétés des collocations

En se basant sur les travaux de Hausmann (1989) et Mel'ćuk (1998), nous distinguons sept propriétés complémentaires définissant la collocation :

1. **l'aspect arbitraire (la non prédictibilité) de l'association lexicale :** même si dans de nombreux cas, la cooccurrence des éléments de la collocation peut être expliquée à l'aide de principes sémantiques, la collocation est définie par Hausmann (1989) comme une association imprévisible et arbitraire lexicalement. Par exemple *prescrire* peut apparaître avec *ordonnance* (dans *prescrire une ordonnance*), tandis qu'il est nettement moins naturel (correct) de dire *prescrire une lettre*.
2. **la fréquence de la cooccurrence des éléments :** cette propriété est le plus souvent nécessaire mais non suffisante de la collocation (Bartsch

6. « *[...] l'occurrence de deux items dans un contexte dans un environnement spécialisé. Une collocation significative est une collocation régulière entre deux items qui cooccurrent ensemble plus souvent que leurs fréquences d'apparition respectives et la longueur du texte dans lequel ils apparaissent pourrait la prédire.* »

2004). Des seuils ont habituellement été utilisés dans la littérature permettant d'éviter d'une certaine façon l'arbitraire ou les effets de style. Nous tenons à mentionner l'absence de ce critère statistique chez Hausmann (1989) et Mel'ćuk (1998).

3. **la transparence sémantique de la collocation :** d'après Cruse (1986) et Hausmann (1989), le sens d'une collocation doit être facilement déductible par un locuteur natif. *Prescrire une ordonnance* est compréhensible même pour un locuteur étranger qui serait toutefois incapable de la produire spontanément.
4. **le caractère binaire de la collocation :** contrairement à Cruse (1986) ou Bartsch (2004), pour Hausmann (1989) comme pour Mel'ćuk (1998), la collocation est essentiellement constituée de deux mots ou de deux lexies, l'un appelé base, l'autre collocatif. Dans des exemples comme *avoir une peur bleue* ou *avoir la tête en l'air* il paraît plus pertinent de parler d'entité ou de constituant plutôt que de lexies. Dans certains cas, il est possible d'avoir affaire à des collocations distinctes mais superposées comme *avoir peur* et *peur bleue* qui sont fusionnées dans *avoir une peur bleue* (Tutin 2008).
5. **la dissymétrie des sens des composants de la collocation :** pour Hausmann (1989), la base de la collocation garde son sens habituel dans le contexte, elle est autonome. Le collocatif, quant à lui, dépend de la base. Ceci peut être remarqué dans les dictionnaires de collocations dans lesquels les collocations sont présentées en fonction de leurs bases (ex. pour *tête*, *faire la tête*, *prendre la tête*).
6. **la notion de sélection lexicale :** Hausmann (1989) souligne que la sélection du collocatif est imposée par la base. Par exemple dans la phrase *J'ai une peur bleue*, le choix de *bleue* n'est pas libre mais imposé par *peur*. Mel'ćuk (1998) explique plus clairement que lorsque l'on veut produire une collocation, le choix du collocatif n'est pas libre mais imposé par la base. Hausmann (1989) et Mel'ćuk (1998) se sont intéressés aux collocations dans un cadre générique de la langue, en s'appuyant sur des dictionnaires. Notre but consiste à identifier les collocations plutôt qu'à les générer, dans des perspectives de compréhension.
7. **la relation syntaxique entre les deux composantes de la collocation :** Hausmann (1989) et Bartsch (2004) considèrent que les éléments de la collocation doivent entretenir une relation syntaxique. Cela donne une définition contraignante des schémas syntaxiques que peuvent avoir la base et le collocatif, en fonction de la catégorie grammaticale de la collocation. Par exemple *la peur paralyse* est une collocation verbale (où le collocatif *peur* est sujet de la base *paralyse* qui est le verbe) dont la structure syntaxique est *nom + verbe*.

   Les collocations peuvent être classées selon leur catégorie grammaticale. Nous présentons pour chaque catégorie les structures syntaxiques les plus fréquentes accompagnées d'exemples[7]. La base est écrite en caractère gras, l'ordre de la base et du collocatif peut varier :

   — collocation nominale :

7. Ces exemples sont empruntés à Daille (2001).

  - **nom** + adjectif : ***amour*** *platonique* ;
  - **nom** + nom : ***bourreau*** *des cœurs* ;
  - **nom** + verbe : *retirer de l'****argent*** ;
- collocation verbale : **verbe** + adverbe : ***exploiter*** *efficacement* ;
- collocation adjectivale : **adjectif** + adverbe : *sexuellement* ***transmissible*** ;

Ces propriétés sont issues de recherches menées sur les collocations comme étant des structures lexicales ou syntaxiques régulières, en vue de la production dans une langue donnée. Ici, nous nous intéressons plutôt à identifier les collocations en tant que connaissances riches linguistiquement. Dans la présente thèse nous retenons la définition statistique de Sinclair et al. (1970), et nous empruntons certaines propriétés linguistiques pour définir les collocations :

- sur le plan lexical : propriété 4) associée à l'hypothèse que la collocation peut comporter deux lexies (*i.e* composés syntagmatiques) ;
- sur le plan de l'usage : propriétés 2) et 6) ;
- sur le plan syntaxique : propriété 7) associée à l'hypothèse qu'une collocation a une catégorie grammaticale et respecte une structure syntaxique spécifique : la base et le collocatif entretiennent une relation syntaxique directe de type (terme, collocatif).

Les propriétés 1), 3) et 5) n'ont pas pas été retenues car plus ancrées en production de la langue qu'en compréhension, et elles traitent la sémantique que nous n'abordons pas dans nos travaux de recherche.

## 3.7 Conclusion

Nous avons dans ce chapitre défini la notion de contextes riches en connaissances, qui sont le pilier de cette thèse. Nous avons également structuré les liens entre CRC, marqueurs et patrons de connaissances qui ont souvent été confondus. Par ailleurs, nous avons abordé les définitions terminologiques et les limites qui les séparent des CRC initialement introduits dans des perspectives termino-ontologiques. L'objectif de cette thèse est d'étudier l'utilité des CRC plutôt dans un exercice d'aide à la traduction terminologique. Pour cela, nous nous intéressons également à une notion particulièrement appréciée, tant en lexicographie qu'en traduction, qui est la collocation. Nous joignons cette notion à celle des CRC, au sens qu'elles contiennent toutes les deux des connaissances pouvant être utiles en TAO.

# Méthodes d'identification automatique de CRC

4

Sommaire

## 4.1 Introduction

Les travaux qui, directement ou indirectement ont questionné la notion de CRC sont très nombreux, tant en lexicographie ou en terminologie, qu'en ingénierie des connaissances ou en extraction d'information. Parmi ces travaux, nombreux sont ceux qui se sont penchées sur l'identification de définitions sans toutefois déterminer les limites qui les séparent des CRC. En effet, les définitions tirent partie de contextes que l'on peut incontestablement juger de riches en connaissances.

Une tendance générale consiste à considérer comme riche en connaissances tout contexte contenant au moins un terme et un marqueur de relation (Meyer 2001). Or, cette représentation pose de nombreux problèmes. D'une part, l'équivalence entre marqueurs de relations et connaissance pourrait laisser entendre que seuls les contextes qui peuvent être représentés sous forme de relations, relèvent de la connaissance. D'autre part, l'idée même de marqueurs de relations pose question. Les contextes qui comportent deux termes et un marqueur identifié automatiquement et non-polysémique sont difficilement repérables et nécessitent une interprétation humaine en amont. En pratique, cette interprétation, est semée d'embûches (Aussenac-Gilles et Condamines 2009; 2012). Parmi ces difficultés, citons les cas où : l'identification de l'un des termes n'est pas évidente comme pour les termes polylexicaux (par exemple un seul élément peut être repéré) ; le marqueur est polysémique (par exemple *être* peut indiquer l'hyperonymie ou l'état), etc.

L'identification des CRC a le plus souvent été abordée en vue d'extraire des définitions terminologiques. Nous étudions dans ce chapitre l'état de l'art lié au repérage des des segments de textes susceptibles de véhiculer des connaissances riches et utiles dans l'illustration de termes. Nous étudions, tout d'abord, les approches à base de PC, puis les approches supervisées et semi-supervisées, ainsi que les méthodes d'extraction des collocations. Dans chaque partie, nous citerons succinctement différents travaux, puis nous détaillerons plus particulièrement ceux qui nous intéressent, en partant des moins génériques pour aller aux plus génériques.

## 4.2 Approches à base de patrons de connaissances

Les méthodes basées sur les patrons de connaissances ont été adoptées dans plusieurs travaux de la littérature. Si Auger (1997) a consacré ses recherches à repérer des définitions avec des PC lexicaux, Rebeyrolle (2000a) a employé des PC lexico-syntaxiques ainsi que des contraintes liées à la ponctuation. Pour évaluer sa méthode, elle a eu recours à un corpus étiqueté manuellement par les définitions. Celles ci serviront de références pendant l'évaluation. Les travaux de Rebeyrolle (2000a) ont permis d'obtenir une précision comprise entre 17,95 % et 79,19 % et un rappel entre 94,75 % et 100 % selon les PC utilisés. D'autres recherches telles que Muresan et Klavans (2002) ont proposé un outil également basé sur des PC comme *is defined as, is called*, et sur des indices de ponctuation de type *()* et

- -. Cet outil sélectionne tout d'abord des définitions candidates à partir de différents articles médicaux disponibles sur le Web. Ensuite, les définitions complexes sont filtrées grâce à un analyseur grammatical. Pour l'évaluation, les sorties de ce système ont été comparées à un ensemble de textes étalons donnant 86,95 % de précision et 75,47 % de rappel.

**Fujii et Ishikawa (2000) :** proposent une méthode pour extraire des descriptions de termes japonais en assimilant le Web à une encyclopédie. Cette méthode repose principalement sur l'utilisation des PC et l'exploitation du format structurel du document (HTML) : les balises. Par ailleurs, Fujii et Ishikawa (2000) se sont servis d'un modèle de langage accompagné d'un processus de classification non supervisée afin de filtrer les sorties.

| **Japonais** | **Français** |
|---|---|
| X toha Y dearu | X est Y |
| X ha dearu | X est Y |
| Y wo X to-iu | Y est appelé X |
| X wo Y to-sadameru | X est définit par Y |
| Y wo to-yobu | Y est appelé X |

TABLE 4.1 – *Exemples de PC utilisés dans Fujii et Ishikawa (2000)*

Cette méthode consiste à extraire des fragments de pages Web sur la base de leur structure, en plus des PC (ex. table 4.1) qui sont fréquemment utilisés dans l'encyclopédie Japanese Word pour décrire les termes. La démarche de Fujii et Ishikawa (2000) est de localiser tout d'abord les extraits susceptibles de contenir des définitions avec des PC, pour en extraire par la suite les segments de textes à l'aide des patrons contenant des balises HTML (ex. <P>....</P>). Une classification hiérarchique bayésienne (Iwayama et Tokunaga 1995) est appliquée dans une étape finale afin d'éviter la redondance des informations extraites. Les problèmes liés à l'extraction de texte à partir des pages Web ainsi que l'apparition de caractères spéciaux provenant des balises HTML peuvent affecter la grammaticalité des informations extraites. Comme solution, cette méthode utilise un modèle de langage de tri-gramme pour ne retenir que les extraits non bruités.

Cette approche a été testée sur 44 termes issus de plusieurs dictionnaires terminologiques. Seulement 27 ont été associés à des définitions candidates qui ont été comparées plus tard à des données de la collection NACSIS (Kando et al. 1999). Les résultats obtenus montrent l'efficacité de l'utilisation du filtrage en termes de précision qui a été améliorée de 63,5 % à 67,9 %. Les expériences menées montrent que la consultation d'environ deux définitions assure la compréhension du sens du terme en question.

**Saggion (2004) :** propose une méthode qui exploite des ressources externes pour identifier les définitions à partir de données textuelles. L'objectif de l'exercice consiste à répondre aux questions « Qui est X ? » ou « Qu'est ce que X ? » que l'auteur considère comme des requêtes de définitions. La contribution de Saggion (2004) repose principalement sur la

mise en œuvre de la notion de « termes secondaires » afin d'identifier les passages susceptibles de contenir des informations définitoires. Il introduit cette notion comme :

> *"Terms that co-occur with the definiendum (outside the target collection) in definition-bearing passages seem to play an important role for the identification of definitions in the target collection [..] Our methode considers nouns, verbs and adjective as candidate secondary terms."* [1] *(Saggion 2004, p. 1928)*

Saggion (2004) considère les termes secondaires comme un indice marquant les définitions. L'intuition derrière l'apparition de cette notion revient à l'observation suivante : cherchant les définitions du mot *Goth* parmi 217 phrases contenant toutes ce mot, Saggion (2004) a remarqué que le mot *subculture* apparaissait régulièrement dans les définitions du mot *Goth* sur le Web. En examinant les 217 phrases de départ, il s'est avéré que seulement 5 d'entre elles étaient des définitions contenant toutes le mot *subculture*. Citons à titre d'exemple la phrase *The goth is a contemporary subculture found in many countries*. À travers cette observation, l'auteur postule qu'un terme donné et son terme secondaire apparaissent, de façon fréquente, dans les définitions, mais rarement dans les autres contextes non définitoires.

Les définitions peuvent être également extraites à partir de ces passages grâce à des PC. L'auteur a utilisé une liste de 69 PC correspondant à des structures de définitions pour la langue anglaise. L'hypothèse consiste au fait que liste de PC utilisée suffira pour repérer les passages définitoires indépendamment du corpus exploité. Ces PC sont répartis en deux ensembles, 36 PC destinés aux questions générales de type « Qu'est ce que X ? », et 33 pour répondre aux questions plus spécifiques « Qui est X ? ». Des exemples de PC sont présentés dans la table qui suit :

| **PC général** | **PC de personnes** |
|---|---|
| define TERM as | PERSON known for |
| TERM and others | PERSON who was |
| TERM consist of | PERSON a member of |

Table 4.2 – *Exemples des PC utilisés dans (Saggion 2004)*

Trois ressources externes ont été sollicitées afin de déterminer les termes secondaires : WordNet [2], Britannica [3] et le Web. Dans WordNet seuls les hyperonymes du mot X à définir, et les mots les plus fréquents dans son contexte sont retenus. Dans Britannica, les termes secondaires ne sont extraits que si les phrases contiennent une référence explicite de X. Pour les mots provenant d'autres pages Web, la phrase retenue doit contenir un PC (de définition) pour considérer ces mots comme des termes secondaires. La table 4.3 présente des exemples de termes secondaires.

1. « *Les termes qui coocurrent avec le definiendum (en dehors de la collection cible) dans les passages contenant des définitions semblent jouer un rôle important pour l'identification des définitions dans les collecions cibles [..] Notre méthode considère les noms, les verbes et les adjectifs comme termes secondaires candidats.* »

2. `https://wordnet.princeton.edu/`

3. `http://www.britannica.com/`

| Mot à définir | Termes secondaires |
|---|---|
| *Aaron Coplan* | *music, american, composer, classical, spring...* |
| *Golden parachutes* | *plans, stock, executive, comensation, million...* |

TABLE 4.3 – *Exemples de termes secondaires utilisés dans Saggion (2004)*

Dans un premier temps, pour identifier les passages contenant des définitions, Saggion (2004) introduit ses requêtes enrichies par les termes secondaires comme des entrées. Ensuite, les phrases retenues comme définitions contiennent soit un PC, soit le mot à définir avec au moins trois termes secondaires. Dans un second temps, Saggion (2004) s'est intéressé uniquement à la clause principale de la phrase tout en tenant compte de la similarité des phrases extraites dans le but de minimiser la redondance.

| Filtre de définition | 20 passages | 500 passages |
|---|---|---|
| (mot à définir + trois termes secondaires) ou PC | 0,29 | 0,38 |
| ((mot à définir ou alias) + un terme secondaire) ou PC | 0,34 | 0,43 |

TABLE 4.4 – *Effet de la relaxation du filtre de définition sur l'extraction des définitions dans Saggion (2004) en termes de F-score*

Les définitions obtenues automatiquement sont évaluées en termes de rappel et précision par rapport à des références. Ici, nous présentons seulement les valeurs de F-mesure. Le système de Saggion (2004) a obtenu un score de 0,236 (meilleur système 0,555, système moyen 0,192) selon la mesure F-score. La deuxième colonne de la table 4.4 est la performance du filtre de définition sur 20 passages. Nous remarquons que la structure *mot à définir ou alias + un terme secondaire ou PC* est plus générique puisqu'elle permet d'extraire plus de définitions que celle qui se base sur la présence de trois termes secondaires ou un PC.

Un des points faibles abordés par Saggion (2004) est le fait de ne pas avoir considéré les noms propres comme des termes secondaires. L'auteur affirme que cela aurait pu non seulement améliorer les définitions des personnes mais aussi les définitions des objets généraux. Il a également constaté que le traitement de différentes figures de synonymie et d'alias du terme à définir aurait pu permettre d'identifier davantage de définitions comme pour *Alberto Tomba* qui a souvent été défini par *Tomba, a three-time Olympic champion...*

**Barrière (2004) :** considère les PC comme un « outil clé » pour l'identification automatique des CRC. Elle a classé les PC présents dans les définitions du dictionnaire American Heritage First Dictionary (AHFD) selon le type de connaissances qu'elles procurent ainsi que selon la relation sémantique qu'elles expriment. L'auteur a défini trois catégories principales de PC : statiques donnant des contextes qui ne sont pas liés à des événements, dynamiques contenant les relations causales et temporelles, et événementielles introduisant des événements. La table 4.5 illustre ces catégories.

| Catégorie | Type de connaissances | Relation sémantique | Exemple |
|---|---|---|---|
| statique | paradigmatique | hyperonymie | *Le chat **est** un félin.* |
| | | synonymie | *Une bicyclette **est** un vélo.* |
| | composition | partie de | *La France **est une partie de** l'Europe.* |
| dynamique | causalité | résultat | *La fumée **sort lorsqu'**on brûle des feuilles.* |
| | | cause | Tuer **cause** la mort |
| | temporalité | temps | *Chuchoter c'**est quand** on parle très bas.* |
| événementielle | événement | instrument | *Mordre c'est couper **avec** les dents.* |

TABLE 4.5 – *Catégorisation des PC*

Par la suite, Barrière (2004) a analysé la généralité/spécificité des PC récoltés par rapport aux domaines et aux relations sémantiques exprimées dans un corpus traitant du domaine de la plongée sous-marine et comptant environ 1 million de mots. Les tables 4.6 et 4.7 présentent des résultats de l'application d'une amorce[4] sur ce corpus, permettant ainsi l'extraction des PC reliant les termes entre eux. Deux catégories de relation se manifestent : dépendantes du domaine, et indépendantes. Les relations dépendantes d'un domaine donné, y sont exprimées avec des PC que l'on peut qualifier d'« authentiques ». Autrement dit, il suffit de changer de domaine pour que ce même PC indique une relation différente de celle de base. Les relations sémantiques indépendantes du domaine, comme l'hyperonymie ou la méronymie par exemple, sont exprimées de la même façon dans le corpus de la plongée sous-marine que dans le dictionnaire AHFD. Pour ces relations, aucun nouveau PC n'a été obtenu (par rapport à ceux identifiés dans le AHFD).

| Relation sémantique | Exemple |
|---|---|
| part of | *Buoyancy-control, body position and propulsion techniques **are part of** both Cavern and Cave Driver training.* |
| hyperonymy | *An air embolism **is another kind of** decompression illness.* |
| cause | *A lung over-expression injury **caused by** holding your breath while ascend.* |

TABLE 4.6 – *Exemples de PC indépendants du domaine trouvé dans le corpus de la plongée sous-marine*

D'autres relations spécifiques au domaine sont exprimées avec de nou-

4. Technique permettant de déduire les PC entre un couple de termes liés par une relation sémantique.

| Relation sémantique | Exemple |
|---|---|
| emmergency measure | *Pure oxygen **is first aid for** any suspected decompression illness.* |
| symptom | *The most common barotrauma **symptom** a driver experience may be mild discomfort to intense pain in the sinus or middle ear.* |
| risk prevention | *Keeping your SPG and high-pressure hose clipped to your left-hand side significantly reduce **the risk of** gauge damage entanglement.* |

TABLE 4.7 – *Exemples de PC dépendants du domaine de la plongée sous-marine*

veaux PC illustrés par la table 4.7. Par exemple, la relation *risk prevention* est spécifique au domaine de la plongée sous-marine dans lequel le PC explicitant cette relation apparaît exclusivement. Pour affiner l'identification des relations, Barrière (2004) a enrichi la structure lexico-syntaxique des PC avec une nouvelle dimension sémantique comme suit :

— l'information lexicale est introduite explicitement par le PC (sa forme lexicale) ;
— l'information syntaxique est exprimée avec les parties du discours pour définir un PC ;
— l'information sémantique est exprimée en indiquant les types relations (ex hyperonymie et synonymie) ce qui permettrait de construire des PC génériques.

La table 4.8 présente des exemples de PC fournissant des informations sémantiques proposés par Barrière (2004).

| PC | Lien sémantique | Catégorie sémantique | Instances |
|---|---|---|---|
| is a &home for | synonymy | home | *house, living-place, roof* |
| are ^colour | hyperonym | colour | *brown, blue, green* |
| is a ^time when | hyperonym | time | *period, moment* |
| is an &amount of | synonym | amount | *quantity* |

TABLE 4.8 – *Exemples d'interaction entre l'information lexicale et sémantique dans le PC. (& : synonyme, ^ : hyperonyme)*

| PC | Partie du discours | Relation | Instances |
|---|---|---|---|
| is a *\|a group of | adjectif | group | *large, small, electic* |
| is a tool *\|p | preposition | function | *to, far* |
| is to *\|r make | adverbe | causal | *really, largely, principally* |

TABLE 4.9 – *Exemples d'interaction entre l'information lexicale et syntaxique dans le PC*

L'auteur a utilisé des mesures statistiques telles que la fréquence d'occurrence et la précision afin de comparer les stratégies adoptées qui dé-

finissent les patrons de connaissances. Les résultats sont illustrés par la table 4.10.

| PC | (i) | (ii) | (iii) |
|---|---|---|---|
| serve to | 1 | 0 | 0 % |
| useful for | 1 | 0 | 0 % |
| made to | 2 | 0 | 0 % |
| intended for | 1 | 0 | 0 % |
| design∗ to | 28 | 2 | 7 % |
| used ∗ \| p | 47 | 11 | 23 % |
| Total | 80 | 13 | 16 % |

TABLE 4.10 – *Statistiques sur des différents PC pour la relation « fonction ». (i : occurrences de PC, ii : nombre d'occurrence n'affichant pas de CRC, iii : pourcentage des occurrences n'indiquant pas la « fonction ».*

Dans la table 4.10 le PC *design*∗ [5] correspond à *design, designs* et *designed*. Le dernier PC *used**|*p* [6], où *p* est une préposition, permet la collecte de 47 PC dont 15 *used to*, 9 *used in*, 8 *used by*, 5 *used for*, 4 *used with*, 2 *used as*, 1 *used up*, 1 *used like*, 1 *used on* et 1 *used after*. Bien que les PC syntaxiques et sémantiques soient fréquents dans le corpus, ils s'avèrent moins précis que les PC lexicaux. En effet, 23 % des contextes du PC *used* ∗ |*p* ne sont pas des CRC.

**Malaisé et al. (2004)** se sont appuyés sur les travaux d'Auger (1997) et de Rebeyrolle (2000a) pour définir une liste de marqueurs de relations adaptés à leurs corpus. Ils ont davantage pris en considération des marqueurs liés à la ponctuation. Malaisé et al. (2004) ont tenté de repérer les définitions pour en extraire ultérieurement les relations entre les termes dans la perspective d'aider à la construction d'ontologie. L'évaluation de ce point a soulevé la problématique d'une faible précision quand il s'agit de marqueurs linguistiques de reformulation plutôt que de marqueurs lexicaux métalinguistiques. Cette remarque a également été mentionnée par Rebeyrolle (2000a). Nous détaillons, ci-après, les marqueurs utilisés dans Malaisé et al. (2004) pour détecter les définitions en français :

— **9 marqueurs métalinguistiques** [7] **indépendant du domaine :** appeler, baptiser, définir comme, dénommer, dénoter, désigner, nommer, signifier, vouloir dire ;

— **11 marqueurs métalinguistiques nominaux :** appellation, acception, concept, dénomination, désignation, expression, mot, nom, notion, terme, vocable ; sont à associer à un verbe support parmi : appliquer, donner, employer, prendre, porter, recevoir, référer, renvoyer, réserver, utiliser ;

— **21 marqueurs lexicaux n'étant pas explicitement métalinguistiques, ou ceux de reformulation :** c'est-à-dire, en d'autres termes, soit, à savoir, en quelques sortes, une sorte de, enfin, il s'agit de,

5. L'étoile désigne toute forme du mot précédent.
6. *|p désigne toute préposition.
7. Renferment un lexique métalinguistique.

entendre par, vouloir dire, indiquer, comme, dit, par exemple, autrement dit, même chose que, équivaloir à, employer pour, marque, expliquer, préciser ;

— **Les ponctuations :** parenthèses, guillemets et tirets d'incise sont également mentionnés dans la littérature.

L'évaluation de cette méthode s'articule autour de deux parties : autour des énoncés définitoires, et autour des unités lexicales extraites à partir de ces énoncés. Ici, nous présentons la première évaluation réalisée sur deux corpus qui relèvent des domaines de « la petite enfance » et « la diététique ». Un corpus de définitions références a été utilisé. Les résultats de la table 4.11 montrent que les PC de type métalinguistique 1 et 2 (indépendant du domaine, et nominaux) donnent une meilleure précision malgré leur faible rappel par rapports aux autres PC utilisés.

| Type | # PC | # PC extraits | # définitions | Précision |
|---|---|---|---|---|
| Méta1 | 7 | 32 | 16 | 50 % |
| Méta2 | 20 | 14 | 7 | 50 % |
| Ling | 24 | 97 | 38 | 39 % |
| Ponct | 4 | 79 | 22 | 27 % |
| **Total** | 55 | 222 | 83 | 37 % |

TABLE 4.11 – *Détection de définitions dans le corpus « petite enfance ».*

Afin de repérer les énoncés définitoires, les travaux présentés dans cette section ont tous utilisé des PC (principalement lexico-syntaxiques) signalant des énoncés riches en informations sémantiques. Par exemple, Rebeyrolle (2000a) et Barrière (2004) ont étudié les structures linguistiques exprimant la définition dans les textes. D'autres travaux comme celui de Saggion (2004) ont choisi d'exploiter les termes comme étant un indice de définition. Cette notion de termes secondaires étant, en pratique, proche à celle de la collocation (mais limitée aux énoncés définitoires) permet d'identifier les contextes définitoires « typiques ».

L'idée première de l'identification des CRC vise la construction des définitions. Or, les travaux qui exploitent les PC pour repérer les CRC ont mis en évidence plusieurs écueils : les PC peuvent être très efficaces dans un texte et très peu dans un autre. En effet, les définitions peuvent être exprimées de manières différentes en fonction du domaine du texte utilisé et de son type du discours. Ainsi, malgré leur large exploitation dans l'état de l'art, il n'existe aucune liste exhaustive qui contiendrait tous les PC étudiés. C'est pour ces raisons que plusieurs recherches se sont orientées vers des méthodes moins dépendantes des patrons de connaissances ou de leur acquisition.

## 4.3 Approches supervisées

Les approches supervisées reposent sur une approche générale référentielle dans laquelle les CRC ont été préalablement identifiés. Ces CRC

de références sont par la suite exploités afin en déduire les traits (i.e critères) qui permettent d'entraîner un classifieur et prédire si un contexte est un riche en connaissances. Fahmi et Bouma (2006) ont proposé une méthode reposant sur l'apprentissage supervisé afin de repérer les définitions dans un corpus allemand du domaine médical. Ce corpus est constitué de pages de Wikipedia. Dans une partie du corpus, ils ont commencé par extraire toutes les phrases contenant le marqueur *to be* afin d'obtenir des définitions candidates. Parmi ces phrases, les définitions ont été repérées manuellement. Ensuite, Fahmi et Bouma (2006) ont déduit les traits permettant de distinguer ces définitions des autres phrases dans cette partie de corpus. Il s'agit de la position de la phrase dans le document, la distribution des mots et des bigrammes, ainsi que des traits syntaxiques comme le type du déterminant et la position du sujet dans la phrase. Ils ont alors intégré ces traits dans trois systèmes d'apprentissage différents : Bayésien naïf, SVM (Support Vector Machine) et MaxEnt (Maximum Entropy) pour identifier les définitions dans les autres parties du corpus qui n'ont pas servi à l'apprentissage. Les résultats obtenus varient de 77 % à 92,3 % (le meilleur fourni par MaxEnt) en termes de précision. Quelques années plus tard, Westerhout (2009), inspiré par Fahmi et Bouma (2006), a proposé une méthode hybride dans laquelle il a eu recours à l'apprentissage supervisé et aux patrons de connaissances. L'auteur a ajouté le type des noms (communs ou propres) et la structure du document aux traits de son système. Les meilleurs résultats (F-mesure = 0,63) proviennent du patron *is a*.

**Kilgarriff et al. (2008) :** se sont intéressés à un autre type de contextes qui ne sont pas des CRC au sens terminologique, mais qui contiennent des informations syntagmatiques, distributionnelles, paradigmatiques, pragmatiques ou autres. Il s'agit d'exemples lexicographiques qui peuvent être perçus comme utiles dans l'illustration des termes. Les auteurs ont développé GDEX (Good Dictionary Examples), un outil qui propose aux lexicographes plusieurs exemples permettant d'illustrer un mot donné. Ils se sont inspirés de la pratique des lexicographes cherchant un bon exemple pour un terme donné. Dans ce travail, ils reprennent les critères de Atkins et Rundell (2008) pour qualifier un bon exemple de « lisible » et « informatif » à l'aide des traits ci dessous :

| **Trait privilégié** | **Trait pénalisé** |
|---|---|
| - longueur de phrase<br>- présence de la collocation souhaitée dans la clause principale<br>- phrase entière<br>- présence d'un troisième collocatif<br>- position de la collocation dans la phrase | - fréquence des mots<br>- présence de pronoms et d'anaphores |

TABLE 4.12 – *Catégories des traits proposés*

Nous classons ces traits selon « lisibles » ou « informatifs » :

— critères informatifs :

1. fréquence des mots : les mots rares (qui apparaissent dans un dictionnaire de mots rares) rendent plus ambiguë la phrase ;
2. présence de pronoms et d'anaphores : ces mots présentent une ambiguïté au niveau du contexte. La phrase sera donc pénalisée ;
3. présence de la collocation souhaitée dans la clause principale : si le couple (base, collocatif) apparaît dans la clause principale de la phrase, celle-ci est privilégiée ;
4. présence d'un troisième collocatif : la phrase contenant un troisième collocatif, et un mot important en plus du couple (base, collocatif) est privilégiée ;
5. position de la collocation dans la phrase : Kilgarriff et al. (2008) ont constaté que les bons exemples introduisent souvent un contexte, ensuite la collocation. Les phrases dont la collocation se situe vers la fin, sont privilégiées.

— critères lisibles :

6. phrase entière : une phrase commençant par une majuscule et finissant par un point est privilégiée ;
7. longueur de phrase : choisir les phrases d'une longueur comprise entre 10 et 25 mots. Celles qui sont très courtes ou longues seront pénalisées.

Dans le but de bien pondérer les traits proposés, Kilgarriff et al. (2008) ont demandé à deux étudiants de choisir les meilleurs exemples illustrant 1 000 collocations, à partir du corpus BNC [8] qui a été remplacé plus tard par le UKWaC. Dans un premier temps, en partant de ces « bons exemples » Kilgarriff et al. (2008) ont déduit grâce à des classifieurs (*cf.* table 4.13) la combinaison adéquat de poids associant le meilleur score à ceux-ci. Dans un second temps, ils ont évalué indépendamment, un par un, ces classiffieurs et les ont ordonné en fonction de leur pertinence (ex. pertinent, régulier...). Pour ce faire, ils ont effectué des jeux de tests basés sur des comparaisons avec le corpus d'apprentissage considéré comme un corpus de référence. D'après les résultats obtenus, la longueur de la phrase et la fréquence des mots sont les traits qui influencent principalement le choix des exemples.

**Didakowski et al. (2012) :** présentent une méthode d'extraction d'exemples à partir d'un corpus allemand. Les 10 meilleurs exemples candidats seront intégrés par des lexicographes dans un dictionnaire numérique. Dans ce travail, la sélection des exemples repose sur les notions de lisibilité et de complexité de la phrase ainsi que l'usage des mots (en anglais, *word usage*). Inspirés par Strauss et al. (1989) et Kilgarriff et al. (2008), Didakowski et al. (2012) définissent des critères qui permettent de qualifier une phrase de bon exemple et qui doivent être, en même temps, opérationnels pour la mise en œuvre de la méthode. Strauss et al. (1989) mentionnent quatre critères de base décrivant un bon exemple. Il s'agit d'une phrase qui :

8. http://natcorp.ox.ac.uk

| **Type** | **Classifieur** | **Trait étudié** | **P** |
|---|---|---|---|
| Weak classifiers | word length | word length | 0,653 |
| | conjunction | conjunctions | 0,644 |
| | sentence length | sentence length | 0,642 |
| | collocations avg | collocation average | 0,619 |
| | Regexp blacklist | capital letters | 0,593 |
| | minimal occurrence | rare words | 0,578 |
| | sences | multisence words | 0,574 |
| | low frequency | low frequency words | 0,572 |
| | position | keyword position | 0,522 |
| | Regexp blacklist | mixed symbols | 0,500 |
| | Regexp blacklist | almost never symbols | 0,652 |
| | verbs | verbs score | 0,489 |
| | wordlist | anaphors | 0,473 |
| | sentence | sentence completeness | 0,472 |
| | high frequency | high frequency words | 0,460 |
| | Regexp blacklist | banned symbols | 0,459 |
| | wordlist minimum | 100 commonest words | 0,457 |
| | Regexp blacklist | blacklist | 0,455 |
| Strong classifiers | weighted score | best result | 0,722 |
| | progressive score | best result | 0,714 |
| | approximate p.s | best result | 0,688 |
| Random result | | | 0,454 |
| Best possible benchmark | | | 0,945 |

TABLE 4.13 – *Résultats des jeux de tests de Kilgarriff et al. (2008)*

1. illustre une description de l'objet ou de l'activité que signifie le terme à définir ;
2. contient les mots qui co-occurrent « typiquement » avec le terme à définir ;
3. est authentique ;
4. contient les mots en relation lexicale/sémantique avec le terme à définir.

Pourtant, ces critères ne peuvent pas tous être appliqués de façon opérationnelle. En effet, le critère 1. est très vague et ne peut pas être automatisé, le critère 3. est une condition nécessaire mais pas un critère de sélection. En revanche, les critères 2. et 4. sont plutôt applicables et auront un rôle principal pendant l'évaluation. Didakowski et al. (2012) ont ajouté d'autres critères afin d'affiner la description du bon exemple :

5. un exemple doit être une phrase complète, bien formée et non complexe ;
6. la phrase doit être indépendante c'est-à-dire compréhensible sans avoir besoin d'un contexte plus large ;
7. le terme visé ne doit pas être un nom propre ;
8. l'ensemble des bons exemples doivent couvrir tous les sens du terme en question.

Les critères 5 et 6 sont également mentionnés dans Kilgarriff et al. (2008). Quant au critère 8, il est particulièrement important puisque ces exemples sont censés être le seul moyen qui précise le sens du terme visé dans le corpus. Les phrases ne correspondant pas aux critères 5 et 7 sont rejetées, alors que les critères 2, 4 et 6. influent seulement sur la qualité du score associé à chaque exemple candidat. Le critère 8 n'est pas appliqué sur les phrases une par une mais plutôt sur l'ensemble des exemples candidats.

Pour mettre en œuvre les notions de lisibilité, complexité et l'utilisation des mots, Didakowski et al. (2012) ont utilisé des outils comme :

— TAGH (Karlsson, 2010) : un analyseur morphologique de l'allemand ;
— Moot (Jurish, 2003) : un étiqueteur de partie-de-discours (POS) ;
— SynCoP (Didakowski, 2008) : un analyseur de dépendances.

L'analyseur SynCop est censé mesurer la complexité et la lisibilité de la phrase. Par exemple, si l'analyseur associe un mauvais score à une phrase, elle est considérée comme non grammaticale ou très complexe.

Didakowski et al. (2012) ont intégré dans leur système les traits « légers » :

1. longueur de la phrase : un bon exemple doit contenir entre 10 et 25 mots ;

2. complétude : la majuscule au début de la phrase avec un signe de ponctuation à la fin, sont un indice viable de la complétude de la phrase
3. tous les mots de la phrases doivent être reconnus par l'analyseur morphologique ;
4. absence de pronoms : la phrase ne doit pas contenir de pronoms ;
5. complexité : la phrase ne doit pas être rejetée par l'analyseur.

Par ailleurs les auteurs privilégient des genres de documents comme les articles scientifiques puisqu'ils sont plus structurés que les journaux de la presse écrite par exemple. D'autres critères dits « souples » ont été utilisés dans le but de trier les exemples candidats retenus (qui correspondent aux traits légers). Nous citons ces critères dans l'ordre de leur importance :

1. les mots de l'exemple doivent faire partie des 17 000 mots les plus fréquents du corpus ;
2. les mots de l'exemple ne doivent pas dépasser 15 caractères ;
3. le mot à définir doit être dans la clause principale de la phrase.

Les 20 premiers exemples correspondant aux critères légers et souples sont retenus. Didakowski et al. (2012) ont évalué leur méthode sur un corpus de 100 millions de mots contenant des textes allemand du 20ème siècle dans les domaines de journalisme, littérature, scientifique ainsi que des transcriptions de paroles. Ce corpus était enrichi par d'autres journaux allemands pour atteindre au final environ un billion de mots.

Après avoir intégré les traits précédents (souples et légers) dans un classifieur, 19 000 phrases sont considérées comme de bons exemples associés à 5 076 mots à définir. En moyenne, chaque mot à définir est associés à 3,7 exemples. Seulement 34 mots n'ont pas eu d'exemples acceptables. Les exemples fournis sont étiquetés comme suit :

1. les exemples qui sont grammaticalement corrects et au moins acceptables ;
2. les exemples qui sont acceptables mais qui demandent des corrections (non majeures) ;
3. les exemples qui ne sont pas acceptables parce qu'ils sont malformés : le mot à définir est utilisé comme un nom propre par exemple.

| **Classe** | **# exemples** | **Précision** |
|---|---|---|
| 1 | 1 8113 | 95,3 % |
| 2 | 342 | 1,8 % |
| 3 | 543 | 2,9 % |

TABLE 4.14 – *Les classes des exemples identifiés dans Didakowski et al. (2012).*

Il est à noter qu'en se basant sur le critère de Strauss et al. (1989), le critère 4 et le critère supplémentaire 8 pour évaluer la qualité des exemples, l'auteur a remarqué que plusieurs exemples contiennent des mots qui sont

liés sémantiquement tels que les co-hyponymes principalement exprimés en forme de conjonction. Il a également remarqué que le critère le plus important d'un exemple est la « structure sémantique » du mot à définir, c'est-à-dire la présence d'une relation sémantique avec un autre mot, sans parler explicitement de marqueurs de relations.

Les travaux de Didakowski et al. (2012) et Kilgarriff et al. (2008) se sont intéressés à l'extraction d'exemples. Ils se sont appuyés sur les notions de lisibilité et de complexité afin de qualifier un contexte de « bon exemple ». Même s'ils présentent des similitudes, ces deux travaux diffèrent sur le fait que dans Kilgarriff et al. (2008), les critères représentent des contraintes qui doivent être respectées par un bon exemple, alors que dans Didakowski et al. (2012), ces traits sont exploités dans le but d'associer des scores aux phrases. Les résultats de Didakowski et al. (2012) sont cohérents avec ceux de Kilgarriff et al. (2008) avec une amélioration considérable en terme de la qualité d'évaluation dans laquelle l'auteur a affiné le type des contextes obtenus (classe et précision).

## 4.4 Approches semi-supervisées

**Navigli et Velardi (2010) :** considèrent comme définition toute phrase pouvant être associée à un automate préalablement généré à partir d'une définition. Cet automate correspond à un PC plus générique que les PC syntaxiques, et s'intéresse plutôt à la structure de la phrase basée sur la position du terme et celle de son hyperonyme. Navigli et Velardi (2010) disposent d'un corpus de définitions extraites de Wikipedia et étiquetées manuellement permettant d'en déduire des modèles de définitions sous forme d'automates finis déterministes. La position des hyperonymes dans les définitions est indiqué grâce à une étiquette préalablement définie par des annotateurs. Ces automates renseignent sur la position de l'hyperonyme du terme à définir dans la phrase.

Dans le corpus des définitions, les phrases respectent une structure formelle bien définie d'une définition. Par exemple, la phrase suivante *In computer science, a graph is a data structure*, correspond à la structure suivante (Storrer et Wellinghoff 2006) :

— clause du terme à définir (CT) contenant le terme à définir et éventuellement son modifieur (*In computer sciences, a graph*) ;

— clause définitoire (CV) contient le verbe utilisé pour introduire la définition (*is*) ;

— clause du défini (CD) contient le genre du terme à définir (souvent l'hyperonyme, *a data structure*) ;

— clause complémentaire (CC) permet de fournir une information supplémentaire sur le terme à définir (n'existe pas dans cet exemple).

Ainsi, la phrase *In computer science, a graph is a data structure*, sera étiquetée par *[In computer science, a graph]CT [is]CV [a data structure]CD []CC*.

Un processus préliminaire consiste à déterminer pour chaque phrase du corpus d'apprentissage son patron, comme suit :

Dans le corpus d'apprentissage, toutes les phrases sont manuellement étiquetées avec CT, CV, CD, CC. La première étape consiste à déterminer tout d'abord l'ensemble *F* des mots fréquents du corpus (ex. in, a, is... ). Ensuite, les termes à définir sont remplacés par une étiquette générique (ex. <*Cible*>). Puis, si un mot appartient à *F*, il reste tel quel, sinon il est remplacé par sa catégorie grammaticale comme *computer* qui sera remplacé par NN dans l'exemple précédent. Ainsi, on génère un premier patron « In NN NN, a $<$ *Cible* $>$ is a NN NN » associé à la phrase de l'exemple *In computer science, a graph is a data structure*.

L'algorithme permettant de construire les automates, est appelé Graphe-1 et contient trois étapes :

1. remplacer tous les mots n'appartenant pas à *F* par . Ainsi l'automate de l'exemple précédent correspondra au patron *en *, une < Cible > est une **. Ce patron exprimé est appelé patron-étoile ;
2. classifier les phrases en mettant dans la même classe celles qui sont associées au même patron-étoile ;
3. construire itérativement les automates génériques (en fonction des parties du discours) correspondant à chaque classe.

Navigli et Velardi (2010) proposent également un second modèle appelé Graphe-2 plus générique que Graphe-1. Graphe-2 est obtenu seulement à partir des différentes clauses (i.e CT,CD et CV) de la phrase. À ce niveau, la clause CC est écartée afin de réduire la variabilité des automates.

| **Algorithme** | **P (%)** | **R (%)** | **F (%)** |
|---|---|---|---|
| Graphe-1 | 99,88 | 42,09 | 59,22 |
| Graphe-2 | 98,81 | 60,74 | 75,23 |
| patron-étoile | 86,74 | 66,14 | 75,05 |
| bigrammes | 66,70 | 82,70 | 73,84 |

TABLE 4.15 – *Évaluation du système de Navigli et Velardi (2010) sur les données de Wikipédia*

Pour mesurer la pertinence de ses systèmes, Navigli et Velardi (2010) proposent une comparaison de Graphe-1 et Graphe-2 avec un simple classifieur de « patron-étoile », permettant d'avoir une première idée de la structure de la définition. Ils ont également comparé ces méthodes avec un système de bigrammes s'appuyant sur un seuil de probabilité pour déterminer si une phrase est une définition. L'évaluation a été réalisée en termes de Rappel, Précision et F-mesure.

La table 4.15 montre une précision très importante concernant les méthodes Graphe-1, Graphe-2 (environ 99 %) et « patron-étoile » (86 %). En revanche, les méthodes à base des bigrammes et du « patron-étoile » ont le meilleur rappel. Le rappel de la méthode Graphe-1 peut être justifié par le défaut de généralité de la méthode par rapport à Graphe-2 et aux autres méthodes. Selon la F-mesure, Graphe-2 semble être le mieux adapté pour identifier les définitions.

**Reiplinger et al. (2012) :** pour faire face au problème de portabilité des patrons, les auteurs ont proposé, eux aussi, une méthode semi-supervisée afin d'extraire des PC reliant des couples de termes liés par une relation sémantique. Ensuite, ces PC ont été exploités dans le but d'extraire des définitions à partir des articles ACL Anthology Reference Corpus (ACL ARC) (Bird et al. 2008). À partir d'un ensemble limité de paires de terme-définition et des patrons définis auparavant, le système de Reiplinger et al. (2012) a acquis de nouvelles paires terme-définition ainsi que de nouveaux patrons de connaissances. Les résultats obtenus montrent que cette technique peut être appliquée pour extraire des définitions.

Des nombreux travaux tels que Saggion (2004), Kilgarriff et al. (2008), Didakowski et al. (2012) interrogent la notion de collocations comme indice de connaissances. Pour cela, nous présentons brièvement les méthodes d'extraction de collocations.

## 4.5 Extraction automatique des collocations

Afin de faciliter le travail des terminologues et des traducteurs, des méthodes de repérage automatique des collocations ont été mises en œuvre. Ces méthodes essentiellement statistiques peuvent être également combinées à des analyses linguistiques telles que l'analyse syntaxique ou morphosyntaxique (Fellbaum 1998, Seretan et al. 2008, Seretan 2011). Ces analyses, qui jouent un rôle de filtre, permettent d'affiner la qualité des collocations obtenues et de les classer selon leurs catégories grammaticales.

### 4.5.1 Systèmes d'extraction automatique des collocations

Les systèmes d'extraction automatique de collocations suivent globalement la même procédure :

1. dans un premier temps, ils identifient les collocations candidates qui correspondent à des paires de mots (*i.e* associations). Certains critères comme le schéma syntaxique peuvent être appliqués comme filtres servant à cibler uniquement les associations les plus intéressantes et ne retenir que les collocations candidates.
2. dans un second temps, ils appliquent leur propre mesure (*cf.* section 4.5.2) pour associer à chaque collocation candidate un score de fiabilité.

**La première étape** est très importante pour la performance de l'extraction des collocations. Elle permet de choisir initialement les collocations candidates, ce qui a une influence considérable sur les résultats de la deuxième étape. Optionnellement, un filtre sur un critère linguistique (*ex.* schéma syntaxique) peut être appliqué sur les collocations candidates afin d'exclure certaines combinaisons jugées moins privilégiées (Nerima et al. 2006, Seretan et al. 2008, Seretan 2011). Parmi ces combinaisons, celles qui incluent des articles, prépositions, conjonctions, auxiliaires, etc. Pour

éviter certains risques, comme l'explosion combinatoire dans des corpus de taille importante, les systèmes d'extraction ne retiennent pas toutes les combinaisons possibles de mots comme candidates, dans l'étape 1. L'espace des combinaisons est ainsi réduit à une fenêtre de mots de dimension prédéterminée (classiquement 5 mots) (Nerima et al. 2006).

**La deuxième étape** se charge d'ordonner les collocations candidates choisies dans l'étape précédente selon le score de la mesure utilisée. Le résultat représentera la liste finale des collocations.

### 4.5.2 Mesures d'extraction de collocations

Les méthodes d'extraction de collocations à partir de données textuelles exploitent principalement leurs propriétés distributionnelles. Parmi les nombreuses mesures qui ont été utilisées, nous retenons :

— **la fréquence des unités textuelles :** cette mesure se limite à la distribution habituelle de l'association entre la base et le collocatif. Elle consiste à compter simplement le nombre de fois où la base et son collocatif apparaissent ensemble dans un corpus donné. Les associations fréquentes sont considérées comme des collocations candidates. La limite de cette mesure réside dans l'identification des éléments à compter : les « collocations » les plus fréquentes ont souvent peu d'intérêt ou ne sont pas significatives ;

— **l'information mutuelle (Fano 1961) :** cette mesure compare la probabilité d'observer la base et son collocatif ensemble (probabilité de la dépendance), avec la probabilité d'observer ces deux éléments séparément (probabilité de l'indépendance). L'inconvénient de cette mesure concerne la nature des associations interprétées qui ne sont pas toujours collocative, notamment des associations entre unités lexicales sémantiquement apparentées telles que *hôpital* et *docteur* ou *maladie* et *patient* qui appartiennent au même champ sémantique ;

— **le Z-score (Berry-Rogghe 1973) :** l'auteur s'est appuyé sur la définition opératoire de Halliday :

> *"The syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at n removes (a distance of n lexical items) from an item x, the items a, b, c [...]"* [9] (Halliday 1961, p. 276)

Cette mesure permet donc de fournir, pour une unité lexicale donnée U, un ensemble ordonné de ses cooccurrents significatifs. Elle mesure la différence entre les fréquences observées pour chaque association formée à partir de U, et les fréquences attendues sous l'hypothèse du hasard. Plus le score d'une association est élevé, plus elle est considérée comme significative.

9. *« L'association syntagmatique des éléments lexicaux, quantifiables, textuellement, comme la probabilité qu'il y aura à n suppression (sur une distance de n éléments lexicaux) à partir d'un élément x, les items a,b,c [...]. »*

Contrairement aux PC, les méthodes d'extraction de collocations se basent principalement sur la distribution des termes dans les corpus, quelque soit le domaine et le genre étudiés. Les collocations représentent ainsi une solution générique pour palier les problèmes de l'utilisation des PC, et fournir d'autres types de connaissances linguistiques.

### 4.5.3 Problèmes de collocations

De nombreux travaux (Kilgarriff 1996, Pearce 2002, Evert 2005) ont réalisé des études comparatives sur la performance des méthodes d'extraction automatique des collocations. Il résulte de ces travaux que la qualité des collocations obtenues peut varier en fonction de plusieurs paramètres tels que : la langue étudiée, la taille du corpus utilisé, ou encore le schéma syntaxique de la collocation. Cependant, l'extraction automatique des collocations n'est pas encore parfaite. Le choix initial des collocations est une question importante qui nécessite la mise en œuvre de critères linguistiques pour distinguer les collocations des termes complexes et d'autres expressions polylexicales (Nerima et al. 2006). Toutefois, les systèmes d'extraction de collocations basés sur des méthodes statistiques ne permettent pas actuellement d'aboutir à une distinction automatique et nette, entre les différentes sous-classes d'expressions polylexicales (Nerima et al. 2006, Seretan et al. 2008) qui constituent généralement un continuum (Wehrli 2000, McKeown et Radev 2000). En effet, la plupart des critères linguistiques mis en œuvre se contentent de fournir une valeur continue sans déterminer le périmètre séparant les collocations des termes complexes. Néanmoins, des travaux prometteurs récents tels que Bride et al. (2015) commencent à proposer des modélisations plus sophistiquées des collocations et ainsi améliorer la qualité des résultats obtenus.

## 4.6 Conclusion

Dans ce chapitre, nous avons illustré l'état de l'art portant sur l'identification automatique (ou semi-automatique) des contextes qui peuvent être qualifiés de riches en connaissances. Deux principaux types d'approches se distinguent : par règles ou par apprentissage. Barrière (2004), par exemple, s'est explicitement donnée pour objectif la description de PC signalant des énoncés riches en connaissances. D'autres méthodes ont essayé de trouver un compromis entre les deux, c'est-à-dire soit en traduisant la structure linguistique en modèle générique (Navigli et Velardi 2010), soit en incluant comme traits les PC. Nous tenons à noter que la plupart des travaux reposent sur la présence du terme à illustrer ainsi que les PC pour identifier les CRC. D'autres, comme Kilgarriff et al. (2008) et Didakowski et al. (2012) ont intégré une information linguistique différente représentant un voisinage « typique » des termes : les collocations. Saggion (2004) a proposé une heuristique dérivée des collocations qui est les termes secondaires. Le principal enjeu des méthodes basées sur des

règles est le faible rappel des PC qu'elles appliquent malgré leur précision. Ceux-ci, dépendent la plupart du temps du genre et du domaine du corpus étudié. En revanche, les collocations sont plus fréquentes en corpus spécialisé et peuvent être comme une solution pour palier le problème des PC. Nous envisageons d'exploiter ces deux connaissances (à savoir les PC et les collocations), riches au sens linguistique, dans la perspective d'aider à la traduction spécialisée.

# Deuxième partie

# CRC monolingues

# Extraction unifiée de CRC

# 5

Sommaire

## 5.1 Introduction

Afin de préparer le travail de traduction, le traducteur peut préalablement construire un glossaire minimal ou une liste terminologique du domaine sur lequel porte son projet de traduction. Le point de départ de toute recherche terminologique d'un domaine de spécialité est la constitution d'une base de documents, ou de corpus disponibles sur la toile. Sur ces derniers, seront appliqués des logiciels d'aide à la traduction, notamment des outils d'extraction terminologique bilingues. Ces outils, connaissant un grand succès auprès des traducteurs, se basent habituellement sur les corpus parallèles largement exploités malgré les limites qu'ils ont montrées. D'une part, ces corpus, assemblés à partir de segments de textes traduits, sont rares pour des couples de langue ne faisant pas intervenir l'anglais ; et d'autre part, la fiabilité de la terminologie extraite dépendra de celle du corpus utilisé (ou de la qualité de l'alignement). Pour cela, les corpus comparables, qui sont constitués de textes authentiques et complets ont commencé, ces dernières années, à être exploités comme étant des ressources d'aide à la construction de dictionnaires bilingues (Bowker et Pearson 2002). Ils sont définis par Bowker et Pearson (2002) comme étant des corpus multilingues qui ne sont pas des traductions à proprement parlé, mais qui partagent certaines caractéristiques telles que la période et le thème. De leur coté, les recherches en extraction de terminologies bilingues se sont elles-aussi tournées vers l'exploitation des corpus comparables (Fung et McKeown 1997, Déjean et Gaussier 2002, Bowker et Pearson 2002).

Les principaux travaux en extraction terminologique bilingue à partir de corpus comparables se basent sur l'hypothèse qu'un mot et sa traduction apparaissent régulièrement dans les mêmes environnements lexicaux (Firth 1957). Les approches standard (Fung et McKeown 1997, Rapp 1999) et par similarité inter-langue (Déjean et Gaussier 2002, Daille et Morin 2005), dédiés à l'extraction de lexiques bilingues à partir de corpus comparables, reposent plus particulièrement sur ce principe. En effet, elles permettent, à partir d'un terme à traduire (dans une langue source) d'obtenir une liste ordonnée de traductions candidates (dans une langue cible). Ces traductions sont le plus souvent obtenues en comparant le contexte traduit, en langue cible, du terme source avec l'ensemble des contextes des termes de la langue cible. Les traductions candidates se présentent sous la forme d'une liste plate qui ne fournit aucune information contextuelle structurée permettant de saisir le contexte d'utilisation du terme visé. Par exemple la méthode standard propose *boule*, *goutte* ou *projection* comme traductions candidates du terme *blob* en vulcanologie, comme le montre la figure 5.1.

En pratique, les systèmes d'alignement terminologique en corpus comparables sont largement utilisés en traduction professionnelle. En termes de performance, ces approches de traduction semblent avoir atteint leurs limites, et les recherches les plus récentes se focalisent plutôt sur l'évaluation de ces approches (Laroche et Langlais 2010). En effet, les résultats va-

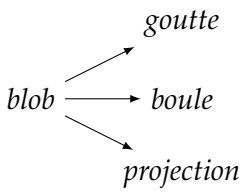


FIGURE 5.1 – *Exemple de problématiques rencontrées en traduction spécialisée : liste de traductions candidates pour blob*

rient selon que le corpus comparable relève d'un domaine général ou d'un domaine de spécialité. Dans le premier cas, la quantité de données disponibles permet de trouver la bonne traduction dans 80 % des cas parmi les 20 premières traductions candidates (Fung et McKeown 1997, Rapp 1999). Dans le deuxième cas, dans lequel les données sont le plus souvent beaucoup moins volumineuses, la précision chute à 60 % pour les 20 premières traductions proposées (Déjean et Gaussier 2002, Morin et al. 2007). Bo et al. (2011) montrent que ces lexiques alignés permettent d'améliorer la qualité d'un système de recherche d'information interlingue. En revanche, lorsqu'il s'agit d'exploiter ces lexiques dans le cadre d'une aide à la traduction, les résultats sont bien plus contrastés (Delpech 2011). La principale difficulté semble être liée à l'absence d'information contextuelle, notamment de CRC. Si l'accès à la terminologie bilingue s'avère indispensable au processus de traduction, le potentiel des méthodes de traduction adoptées doit être amélioré au moyen d'une contextualisation pertinente des termes. En effet, il faut être capable d'appréhender le sens exact d'un terme et de savoir l'employer correctement afin de choisir la bonne traduction parmi *goutte*, *boule* et *projection* (dans l'exemple précédent). Notre objectif consiste à proposer pour chaque terme (ou sa traduction candidate) des CRC aidant à sa traduction.

## 5.2 Compréhension en traduction spécialisée

Un courant important en traductologie s'est penché sur le processus organisationnel déployé par le traducteur lorsqu'il produit une traduction. Une des méthodes expérimentales les plus utilisées repose sur le protocole de verbalisation, appelé aussi raisonnement à voix haute (*think aloud protocol*). Ce protocole consiste à demander au traducteur d'exprimer à voix haute toutes les pensées qui lui viennent à l'esprit dans l'exécution de sa tâche. Pour Durieux (1988), par exemple, toute activité de traduction comprend une étape de lecture, de compréhension et d'écriture. Les connaissances, étant inhérentes à la compréhension du texte, jouent un rôle essentiel dans le travail du traducteur. Pour Daniel (1993), concrètement, l'acquisition de connaissances *ad-hoc* désigne la recherche documentaire dans laquelle le traducteur utilise des outils extérieurs au texte à traduire pour acquérir les connaissances lui permettant de parvenir au niveau de compréhension requis du texte de départ et de ré-exprimer de manière adéquate le contenu de ce texte en langue d'arrivée (Daniel 1993, p. 76). En effet, des cogniticiens comme Van Dijk et al. (1983) expliquent que le lecteur construit diverses représentations à partir de la lecture du

texte et qu'il comprend trois niveaux : la base de texte, les éléments les plus importants du texte et les « modèles de situation » auxquels renvoient les événements relatés dans les textes, ce qui pourraient correspondre dans notre cas aux CRC.

## 5.3 Contexte

Les banques terminologiques et les dictionnaires sont des ressources linguistiques précieuses qui facilitent l'accès aux connaissances des domaines spécialisés. Ces ressources proposent généralement pour un terme à illustrer une définition, des exemples d'utilisation et d'autres termes en relation. Habituellement, le terme à illustrer est considéré comme « terme favori » et les autres termes reflètent des relations paradigmatiques comme la synonymie ou l'hyperonymie. La définition associée au terme favori est souvent de nature encyclopédique et les quelques exemples de contextes proposés, lorsqu'ils existent, ne couvrent qu'une partie des usages de ce terme favori (Bowker 2011). Des récents travaux laissent entendre que les banques terminologiques actuelles n'ont pas connu d'évolution depuis les années 60 (Bowker 2011). Elles contiennent trop peu d'informations contextuelles et les connaissances sur l'usage du terme sont assez limitées (p. ex. des exemples de collocations peuvent être indiqués mais cela n'est pas systématique). En outre, les définitions fournies par les dictionnaires comme les banques terminologiques sont généralement insuffisantes pour permettre la compréhension du terme.

D'après des expériences menées auprès d'apprentis traducteurs, Varantola (1998) détaille les informations que recherchent les traducteurs lorsqu'ils se tournent vers leurs sources de référence. Ces informations se décrivent comme suit :

> *"[...] they often want to know how the expression behaves grammatically and what of lexial, sentence, paragraph or text environment it normally occurs in. At a higher level, they wish to know whether the expression is appropriate for the context, subject field, text type or register in question."* [1] (Varantola 1998, p. 181)

Ces informations renvoient donc à la notion de CRC pour les traducteurs. Une partie de ces informations est accessible à partir des entrées de dictionnaires. Cependant, Varantola (1998, p. 181) indique que les besoins des traducteurs sont en général plus importants qu'un simple accès aux entrées de dictionnaires, et qu'il est nécessaire d'avoir des contextes plus longs sans toutefois en préciser la longueur. Les corpus spécialisés représentent un réservoir conséquent d'informations contextuelles pour analyser le fonctionnement des termes. Cependant, tous les contextes dans lesquels les termes apparaissent ne sont pas utiles à leur compréhension.

1. *« [...] ils souhaitent savoir comment l'expression se comporte grammaticalement et dans quel phrase, paragraphe ou texte syntaxique elle occurrent. À un niveau plus élevé, ils souhaite savoir quelle expression est approriée pour le contexte, le sujet du domaine, le type du texte ou le registre en question. »*

Dans ce cadre, les CRC jouent un rôle prépondérant dans la compréhension des termes et renseignent sur leur fonctionnement en corpus de spécialité.

Dans la perspective d'aider à la traduction terminologique, nous souhaitons proposer pour chaque terme à traduire (ou à sa traduction candidate) des contextes contenant des connaissances conceptuelles et linguistiques qui l'illustrent dans un discours spécialisé. Nous postulons que ces contextes riches du point de vue des linguistes sont également utiles pour les traducteurs. Nous mettons en œuvre deux méthodes.

1. La première méthode s'appuie sur la définition classique des CRC. Elle exige la présence du terme à illustrer ainsi que des PC lexico-syntaxiques afin d'extraire des contextes permettant l'accès à la dimension conceptuelle du terme. Ces PC, intégrant également des termes du domaine, représentent des triplets (Terme1-Relation-Terme2). Il s'agit d'une représentation plus simple des RDF (Resource Description Framework) souvent utilisés pour modéliser les liens entre les connaissances (Lassila et Swick 1999). Les PC exploités ont pour but d'expliciter des relations définitoires du terme, telles que l'hyperonymie et la méronymie. En effet, savoir qu'un terme est en relation avec tel autre terme contribue déjà à le définir. Nous qualifions ces contextes de riches en connaissances conceptuelles (CRCC).
2. La seconde méthode exploite les collocations du terme visé comme un point d'ancrage pour identifier des contextes contenant des connaissances sur le fonctionnement linguistique du terme dans un usage spécialisé. En effet, les collocations peuvent être perçues comme un voisinage typique renseignant sur l'usage du terme en question. Ces contextes sont désormais qualifiés de riches en connaissances linguistiques (CRCL).

Nous exploitons pour notre travail deux corpus comparable spécialisés de discours scientifiques, dans les domaines de la vulcanologie et du cancer du sein, et en deux langues : français et anglais.

## 5.4 Travaux connexes

Parmi les recherches récentes, Schumann (2012b) a entrepris d'extraire des CRC à partir du Web dans le but d'enrichir une banque terminologique en langue russe. Les contextes ont tout d'abord été repérés au moyen de PC, puis ordonnés grâce à une méthode supervisée. Ce travail est similaire au nôtre. Néanmoins, dans le cadre de notre problématique, nous étudions l'identification des CRC dans un corpus spécialisé de taille modeste sans disposer des données nécessaires permettant d'appliquer des méthodes d'apprentissage. Marshman (2014) a étudié la nécessité d'utiliser des ressources terminologiques mettant en évidence des CRC extraits par

des PC, telles que CREATerminal [2]. Ces recherches ont également montré l'utilité des ressources enrichies par des CRC, particulièrement pour des étudiants traducteurs. Une des difficultés majeures dans le domaine des PC tient au fait qu'il n'existe aucune bibliothèque de PC qui manifesterait l'aspect cumulatif de ces travaux. Pour chaque nouvelle étude, il faut refaire une synthèse des études existantes afin d'établir des listes de PC. Par ailleurs, les travaux concernant la variation dans le fonctionnement des PC sont encore récents. En effet, selon Marshman et al. (2008), bien que l'intérêt des PC pour repérer les CRC soit indéniable, leur identification est coûteuse si l'on doit chercher à les réutiliser pour d'autres études, dans d'autres types de corpus. C'est une des raisons pour lesquelles nous envisageons de faire appel aux collocations pour palier les limites des PC.

Les exemples de Kilgarriff sont incontestablement des contextes riches illustrant l'usage du terme. L'objectif de Kilgarriff est de proposer aux lexicographes de nouveaux outils facilitant l'accès aux contextes susceptibles de devenir de bons exemples de dictionnaires lexicographiques. Bien que notre objectif soit similaire à celui de Kilgarriff, nous visons plutôt les contextes qui fournissent des connaissances spécialisées pouvant être utiles pour les traducteurs. Notre hypothèse est celle de la possibilité de sélectionner parmi les occurrences d'un terme, celles où ce terme est employé dans un CRC.

## 5.5 Patrons de Connaissances pour connaissances conceptuelles

Nous présentons dans cette section les relations que nous examinons, le problème de l'instabilité des PC, ainsi que la méthode les mettant en œuvre.

### 5.5.1 Relations examinées

Nous avons mentionné dans le chapitre 4 l'intérêt qu'accordent plusieurs travaux à l'identification des marqueurs de relations dans les définitions en textes spécialisés (Hermans 1989, Flowerdew 1992a;b, Pearson 1998). En effet, savoir qu'un terme est en relation avec tel autre terme contribue, plus ou moins clairement, à le définir. Ces travaux ont considéré trois relations, à tort ou à raison, comme étant universelles que ce soit en terminologie ou en ingénierie des connaissances. Il s'agit des relations d'hyperonymie, de méronymie, et de cause.

La relation d'hyperonymie est connue comme étant la plus structurante, à la fois d'un point de vue linguistique puisqu'elle est fréquemment utilisée pour définir un mot dans une perspective aristotélicienne (Cruse 2002) ; et d'un point de vue de l'ingénierie de connaissances (notamment

2. Interface fournissant des contextes aidant à la traduction terminologique (anglais-français) dans le domaine du cancer du sein.

dans les thésaurus) du fait qu'elle assure l'héritage des propriétés de l'hyperonyme par les termes spécifiques (Hearst 1992). La table 5.1 illustre deux exemples de propriétés transmises par hyperonymie.

| **Terme** | **Hyperonyme** | **Contexte** | **Marqueur de relation** | **Propriété** |
|---|---|---|---|---|
| Tigre | animal | *Le tigre est un animal carnivore* | être_un | carnivore |
| Kilauea | volcan | *Kilauea est un volcan très actif* | être_un | actif |

TABLE 5.1 – *Exemples de connaissances transmises par l'hyperonymie*

Certains travaux visent à identifier des relations de méronymie dans des textes afin d'essayer de proposer des marqueurs de relation, de construire des PC, voire d'étudier la « portabilité » des PC d'un genre textuel à l'autre, c'est-à-dire leur degré de variabilité d'un corpus à l'autre (Barrière 2004, Condamines 2009). La relation de cause a elle aussi été l'objet de nombreuses réflexions depuis la philosophie grecque. Elle intéresse particulièrement la terminologie et l'ingénierie des connaissances qui, travaillant à partir de corpus techniques, rencontrent souvent des contextes qui peuvent être interprétés comme marqueurs de cause. Parmi ces trois relations examinées dans le cadre du projet CRISTAL, nous ne retenons dans la présente thèse que les relations d'hyperonymie et de méronymie. Les marqueurs de causes ne sont pas systématiquement adaptés à être intégrés dans des systèmes informatiques. Leurs mise en œuvre et projection sur les corpus est un exercice complexe (Aussenac-Gilles et Condamines 2009).

### 5.5.2 Stabilité des PC

L'idée première de l'étude des connaissances conceptuelles a été d'associer systématiquement une interprétation de type sémantique à des PC stables. Or, avec un recul de plusieurs années, il apparaît que cette systématicité n'est pas toujours garantie. Plusieurs raisons expliquent ce constat : un même PC peut renvoyer à plusieurs relations, d'où un problème de polysémie (Meyer 2001) ; ou encore, un même PC peut être très efficace dans un texte et très peu dans un autre. Cette instabilité est due au fonctionnement discursif qui dépend du genre et du domaine du discours et qui est fait de polysémie. Cela rend difficiles les traitements visant l'extraction des connaissances. Si l'intérêt des PC pour repérer les CRC est clair, leur identification est coûteuse et l'on doit chercher à les réutiliser pour d'autres études, dans d'autres types de corpus.

Certaines relations très spécifiques au domaine peuvent être identifiées en corpus, sans pour autant être l'objet d'une analyse systématique. Les études portent alors plutôt sur la définition de méthodes pour repérer et interpréter ces relations pour chaque nouveau corpus (Condamines et Rebeyrolle 2000; 2001). Un des objectifs du projet CRISTAL est de construire une bibliothèque de PC pouvant être utilisée dans différents corpus indépendamment du genre et du domaine du texte traité.

### 5.5.3 Méthode

Nous postulons tout d'abord que le contexte d'apparition d'un terme est un CRC candidat. Cependant, les contextes d'apparition d'un terme, qui peuvent être nombreux, ne sont pas tous pertinents. Voici par exemple deux contextes du terme *diabète de type 2* :

a. *...souffrent de diabète, dont le* ***diabète de type 2****...*
b. *Le* ***diabète de type 2*** *est un diabète qui s'accompagne souvent d'un excès de poids.*

Les contextes précédents, correspondent à des structures conceptuelles exprimées par les PC d'hyperonymie *est un* et *dont*.

Nous exprimons ces PC sous la forme d'un triplet *($terme_1$, PC, $terme_2$)* avec $terme_1$ et $terme_2$, deux termes distincts du corpus étudié. La démarche suivie consiste à :

1. **pré-traiter le corpus :** le corpus a d'abord été normalisé et segmenté en occurrences de formes, ensuite étiqueté syntaxiquement avec l'outil Treetagger. Ce dernier a également servi à segmenter le corpus en occurrences de phrases. Le corpus final associe à chaque mot sa partie du discours et son lemme. En outre, lorsque des phrases apparaissent en double dans le corpus, une seule occurrence a été conservée.
2. **extraire la terminologie :** pour chaque langue du corpus comparable, la terminologie a été extraite automatiquement avec l'extracteur terminologique TermSuite (Rocheteau et Daille 2011, Cram et Daille 2016). Cet outil, qui couvre les termes simples et complexes, se base principalement sur des patrons syntaxiques (sous forme d'expressions régulières) pour repérer les termes complexes ; ainsi que sur un score de spécificité (fréquence d'apparition dans le corpus en question par rapport à la fréquence d'apparition dans un corpus général) pour identifier les termes simples.
3. **projeter les PC dans le corpus :** les marqueurs de relations ont été instanciés par le dictionnaire d'expressions régulières du logiciel Caméléon (Séguéla 2001) en associant à chaque marqueur de relation un ensemble de PC correspondant. En pratique, les PC sont des expressions régulières permettant la mise en œuvre informatique des marqueurs de relations. Ils sont projetés dans le corpus après avoir été vérifiés manuellement par des linguistes. Chaque PC est appliqué deux fois, dans les deux sens de la relation. Par exemple, un PC d'hyperonymie est projeté une première fois avec le terme en question en position d'hyperonyme ; puis une deuxième fois en position d'hyponyme. Les PC s'appuient sur la forme lemmatisée des termes. Par ailleurs, les PC ont été exploités seulement dans la forme (active ou passive) dans laquelle ils ont été fournis sans appliquer la forme opposée.

4. **retenir les contextes candidats :** les phrases contenant le PC en question sont des contextes potentiellement riches en connaissances conceptuelles, désormais considérés comme des CRCC candidats.

Cette méthode a été mise en œuvre par l'intermédiaire d'un outil que nous avons développé en Python.

Plusieurs travaux traitant des contextes tels que (Meyer 2001) et (Martínez et al. 2009) ont choisi de travailler sur des unités textuelles plus petites que les phrases. Ces contextes sous-phrastiques risquent de ne pas fournir assez d'informations sur le terme en question. En outre, dans les paragraphes, le risque est plus élevé d'extraire des informations imprécises concernant ce terme. Afin d'éviter cet écueil, nous avons fait le choix de travailler uniquement sur des phrases entières (qui commencent par une majuscule et se terminent par un signe de ponctuation) comme étant un compromis, et ainsi considérer le contexte (a) comme un contexte non-intéressant. Ce critère a également été utilisé par Kilgarriff et al. (2008) et Didakowski et al. (2012).

## 5.6 Collocations pour connaissances linguistiques

Nous présentons la mise en œuvre des collocations pour identifier des connaissances linguistiques et nous discutons la méthode proposée.

### 5.6.1 Méthode

Plusieurs mesures d'association ont été appliquées pour extraire automatiquement des collocations. Si l'Information Mutuelle (Fano 1961) permet d'identifier des unités lexicales qui apparaissent plus souvent ensemble que séparément, le Z-score (Berry-Rogghe 1973) est plutôt privilégié pour déterminer les collocatifs candidats d'un terme donné. Dans ce travail, nous associons à une liste de termes donnés leurs meilleurs collocatifs en nous appuyant sur la mesure du Z-score puisque nous connaissons *a priori* les termes que nous souhaitons illustrer. Ces collocations serviront, par la suite, à sélectionner des contextes potentiellement riches en connaissances linguistiques : des CRCL candidats.

Les mesures d'association peuvent également être combinées à des analyses linguistiques telles que l'analyse syntaxique (Fellbaum 1998). Ces analyses, jouant un rôle de filtre, permettent d'affiner la qualité des collocations obtenues et de les classer selon leurs schémas syntaxiques. Evert et Krenn (2005) montrent la nécessité de distinguer les catégories syntaxiques des collocations avant d'appliquer une mesure d'association. Ici, nous nous appuyons sur la définition des collocations que nous avons adoptée dans la section 3.6. Nous retenons alors trois catégories de collocations nominales dans lesquelles la base est un terme à illustrer et le collocatif est un nom, verbe ou adjectif.

Pour les collocatifs conjugués au participe présent ou au participe passé, ils ont été intégrés dans la catégorie des adjectifs. En effet, établir

une différence entre participe passé et adjectif est difficile pour l'étiquetage automatique en parties de discours (POS). S'agit-il dans l'apposition de la phrase *L'usine sérieusement endommagée par l'explosion, devra être arrachée et entièrement reconstruite* d'un adjectif ou d'une phrase passive inachevée elliptique ? Il semble pour autant justifié d'« admettre un continuum entre les deux valeurs [...] » (Noailly 1999, p.19) d'autant plus qu'il n y a pas de différence de sens. Par conséquent, les limites entre les collocatifs adjectifs ou participes présents sont, elles aussi, fluctuantes. Nous regroupons alors les participes passés et les participes présents dans la même catégorie que les adjectifs.

Guillaume (1984) aborde la grande proximité de l'infinitif relativement à la catégorie du nom puisqu'il peut porter un statut « pleinement » nominal. C'est-à-dire que l'infinitif participe dans le discours de la nature du nom. Nous regroupons alors les noms et les infinitifs dans la même catégorie. Le choix d'intégrer les participes présents et passés dans la catégorie des adjectifs, et les infinitifs dans celle des noms permet de résoudre le problème de conflit d'étiquette morphosyntaxique.

Ainsi, nous obtenons, à la fin de cette étape, trois schémas syntaxiques de collocations dans lesquels la base est un terme connu :

— (**terme**, adjectif ou participe présent/passé) ;

— (**terme**, nom ou verbe à l'infinitif) ;

— (**terme**, verbe fléchi).

Après avoir filtré les mots outils dans le corpus, nous avons repéré, pour chaque terme à illustrer, les collocations constituées de deux mots pleins dans une fenêtre bigramme : un mot avant ou un mot après la base (sans compter les mots vides) en respectant les schémas syntaxiques étudiées. Afin d'extraire les CRCL candidats nous avons suivi les deux étapes suivantes :

1. identifier pour un terme à illustrer ses collocatifs en fonction de leur catégorie syntaxique et les ordonner selon le Z-score ;
2. parcourir les collocatifs de chaque terme à illustrer et retenir les collocations (terme à illustrer, collocatif) qui procurent au moins un contexte phrastique (une phrase entière). Les contextes retenus sont ceux dont les mots pleins contiennent le terme et son collocatif dans une fenêtre de bigramme de mots.

Les CRC candidats de la table 5.2 sont identifiés par des collocations dans lesquelles la base est un terme à illustrer.

### 5.6.2 Discussion

Concrètement, les catégories grammaticales retenues des collocations correspondent à des schémas syntaxiques également exploités en extraction terminologique (Roche 2004) en particulier, pour les termes complexes. Étant en outre des expressions polylexicales, en domaine de spécialité, les collocations telles que définies par Sinclair se rapprochent des

| Terme à illustrer | Collocatif | CRC candidat |
|---|---|---|
| Gaz | carbonique | *Ce **gaz carbonique** qui, transformé par les plantes, a donné de l'oxygène, indispensable à la vie.* |
| Gas | dissolved | ***Gas dissolved** in the molten rock expanded and literally blew the volcano apart...* |
| Cendre | retombée | *Les explosions phréatiques se font plus violentes qu'en 1792, et deux ou trois d'entre elles provoquent des **retombées de cendres** sur les villes du prêcheur.* |
| Cendre | retombée | *Veaucoup d'habitants du prêcheur et de ses environs viennent se réfugier à Saint-Pierre, épargnée par les **retombées de cendres**.* |

TABLE 5.2 – *Exemples de CRC candidats identifiés par des collocations*

termes complexes, notamment les collocations nominales et adjectivales. Ce problème est l'une des difficultés connues rencontrées lors de l'extraction automatique des collocations pour laquelle *roche magmatique* (correspondant à la structure nom + adjectif) fait partie du bruit. La figure 5.2 illustre l'intersection de l'ensemble des collocations avec celui des termes complexes. Cette intersection représente les associations lexicales partageant des critères de syntaxe et de co-occurrence. Toutefois, ces deux notions se distinguent par leurs caractéristiques sémantiques. Dans notre travail, nous ne traitons pas l'aspect sémantique des collocations qui absorbent également une partie des termes complexes extraits par le z-score.

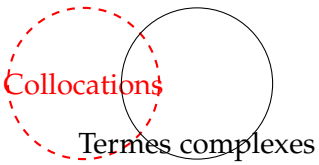


FIGURE 5.2 – *Collocations versus termes complexes*

## 5.7 Conclusion

Les ressources conventionnelles utilisées en traduction terminologique sont encore loin d'être satisfaisantes quand il s'agit d'un terme très technique ou lorsque le traducteur n'est pas expert du domaine. Notre premier objectif dans la présente thèse est de compléter les connaissances qui manquent dans ces ressources, à savoir des contextes authentiques de référence. Pour cela, nous nous sommes appuyés sur la notion de contextes riches en connaissances introduite dans un cadre terminologique. Nous avons enrichie cette notion par une dimension linguistique en qualifiant également les collocations de riches en connaissances. Nous avons ainsi distingué deux types de CRC : CRCC et CRCL extraits en corpus de spécialité. Nous postulons dans la suite que ces CRC peuvent représenter

une information complémentaire aux ressources habituellement utilisées en traduction, et qu'ils peuvent également être utiles pour les traducteurs. Nous évaluons dans le chapitre suivant les CRC candidats fournis par nos méthodes, et nous étudions la validité de notre hypothèse dans un cadre expérimental de traduction.

# Évaluation

6

Sommaire

## 6.1 Introduction

Comme mentionné dans le chapitre 4, plusieurs études ont observé les variations des marqueurs selon les domaines (Bodson 2005) et les genres (Condamines et Rebeyrolle 2000, Condamines 2002, Jacques et Aussenac-Gilles 2006). Ces études comparatives ont ainsi mis en évidence un équilibre, encore difficile à cerner, entre variabilité et stabilité des marqueurs. Les études linguistiques qui se sont penchées sur la portabilité des marqueurs d'une langue à l'autre sont eux beaucoup plus rares hormis la thèse de Marshman (2007). Pearson (2000) s'est intéressée à l'utilisation de corpus alignés (et plus seulement comparables) et a montré que de tels corpus constituaient déjà un apport non négligeable dans le repérage de triplets terme-relation-terme. En effet, lors de la traduction, certains marqueurs ne sont pas traduits ou prennent une forme typographique plutôt que lexico-syntaxique. L'un des objectifs du projet CRISTAL a été d'étudier les relations d'hyperonymie, de méronymie et de cause pour élaborer une base de marqueurs de relations stables en langue source quel que soit le domaine et le genre du texte traité. Dans la présente thèse, nous nous contentons d'exploiter l'hyperonymie et la méronymie pour identifier les CRCC.

La mise en œuvre informatique des marqueurs de relations se fait à travers les PC qui correspondent pratiquement à des automates à états finis ou des expressions régulières balayées par un programme qui cherche à retrouver en corpus la séquence d'éléments ainsi définis. Ces éléments peuvent être des mots, lemmes, catégories syntaxiques, symboles pour marquer les répétitions par exemple, etc. L'application des PC nécessite alors une analyse plus ou moins fouillée et préalable du corpus étudié : analyse syntaxique, lemmatisation, etc. L'informatique fournit de bons instruments de repérage de régularités de forme et à ce titre, les patrons jouent un rôle majeur. Malgré tout, leur mise au point est loin d'être triviale, comme l'ont montré (Rebeyrolle et Tanguy 2000). C'est une tâche complexe qui fait appel à une chaîne de traitements automatiques ou manuels des textes avant de pouvoir projeter des patrons de connaissances.

Outre qu'elles procurent une connaissance linguistique appréciée à la fois par les terminologues et les traducteurs, les collocations peuvent être considérées comme une solution de repli pour palier les limites des PC. Nous avons adopté une définition statistique pour exploiter les collocations. Cela nous évitera l'interprétation sémantique des associations lexicales qui seront identifiées. Nous détaillons dans ce chapitre l'évaluation des CRCC et CRCL extraits en corpus monolingues de spécialité en vue d'aider à la compréhension ainsi que leurs apport dans un environnement de TAO. Nous réalisons deux évaluations dont la première consiste à étudier la richesse linguistique des CRC candidats. La deuxième évaluation vise à valider nos hypothèses et étudier les CRC candidats, plutôt, dans le cadre d'un exercice de traduction.

## 6.2 Ressources

Nous avons appliqué l'extraction des CRC sur les données fournies dans le cadre du projets CRISTAL. Nous détaillons seulement les ressources sur lesquelles ont porté nos expérimentations, et nous décrivons succinctement les recherches menées sur les marqueurs de relations et les PC dans CRISTAL.

### 6.2.1 Corpus comparables

#### Vulcanologie

Ce corpus comparable vulgarisé a été construit par Josselin-Leray (2005). Il est composé de documents scientifiques contenant environ 400 000 mots par langue, obtenus grâce à une recherche thématique à partir de journaux et magazines tels que *Le Monde*, *Sciences et avenir*, *Sciences et Vie* pendant la periode 1980-2002. L'ensemble des documents ont été nettoyés et normalisés à travers les traitements suivants réalisés par la plateforme TermSuite [1] : segmentation en occurrences de formes, étiquetage morphosyntaxique, lemmatisation et extraction terminologique.

#### Cancer du sein

Nous disposons d'un corpus comparable spécialisé portant sur le domaine du cancer du sein (chez la femme) en français et anglais. Ce corpus a été collecté à partir d'articles scienfitiques relevant du domaine médical et provenant du portail Elsevier [2]. Les documents sont retenus sous la contrainte de l'apparition du terme *cancer du sein* (ou *breast cancer*) dans le titre ou dans le résumé qui sont publiés entre 2001 et 2008. Ce corpus compte environ 470 000 mots en français et en anglais.

### 6.2.2 Marqueurs de relations

Bien que la littérature sur les marqueurs de relations soit abondante, il n'existe aucune base de connaissances recensant l'ensemble des PC d'hyperonymie, de méronymie et de cause, ni d'analyse systématique à grande échelle de ces patrons. La contribution de Lefeuvre et Condamines (2015) dans CRISTAL a été de constituer cette base de connaissances et d'analyser chaque candidat-marqueur afin d'en donner une description linguistique fine.

Cette étude s'est déroulée en deux étapes :

1. Élaboration de la liste des candidats-PC en français et en anglais pour les relations d'hyperonymie, de méronymie et de cause. À partir des travaux existants et dans la lignée des travaux (Hearst 1992,

1. http://termsuite.github.io/
2. http://www.elsevier.com

Garcia 1998, Condamines et Rebeyrolle 2000, Séguéla 2001, Cruse 2002, Auger et Barrière 2008, arcalón martínez 2009) une liste a été construite, contenant les marqueurs les plus exhaustifs possibles en français pour trois relations : hyperonymie, méronymie, cause. Pour les marqueurs anglais, une première liste de marqueurs a été dressée à partir d'une étude bibliographique. Cette liste a ensuite été enrichie par la traduction de certains marqueurs de relation français. Une première validation a été effectuée en vérifiant dans le corpus COCA [3] (Davies 2008) les contextes d'apparition des nouveaux candidats-marqueurs anglais ainsi obtenus. La relecture de cette liste par une linguiste anglophone a ensuite permis de valider la liste finale suivante :

| **Marqueurs de relations** | **FR** | **EN** |
|---|---|---|
| Hyperonymie | 33 | 35 |
| Méronymie | 95 | 99 |
| Cause | 192 | 247 |

TABLE 6.1 – *Nombre de candidats-marqueurs par relation et par langue*

2. Analyse des occurrences des candidats-marqueurs dans deux corpus de vulcanologie [4] et d'oncologie (Lefeuvre et Condamines 2015) :

| **Corpus** | **Oncologie (période)** | **Vulcanologie (période)** |
|---|---|---|
| Scientifique | 200 000 mots par langue (2008-2009) | 400 000 mots par langue (1980-2012) |
| Vulgarisé | 200 000 mots par langue (2001-2008) | 400 000 par langue (1980-2002) |

TABLE 6.2 – *Corpus utilisés dans l'étude de la stabilité des marqueurs de relations*

Dans nos travaux de recherche, nous avons retenu les marqueurs les plus productifs et stables d'hyperonymie et de méronymie. Nous avons récupéré cette liste de PC associés à des marqueurs de relations dans le but de les appliquer dans l'extraction des CRCC.

| **Marqueurs d'hyperonymie (FR)** |
|---|
| `X_ETRE_UNE_SORTE_DE_Y` |
| `X_EST_LE_Y_LE_PLUS` |
| `X_ET_AUTRES_Y` |
| `Y_ET_ADVERBE_DE_SPECIFICATION_X` |

TABLE 6.3 – *Exemples de PC exploités pour le français (X est un terme et Y son hyperonyme)*

### 6.2.3 Liste terminologique d'évaluation

Nous avons sollicité des experts des corpus exploités (vulcanologie et cancer du sein) afin de repérer une liste de termes pouvant être difficiles à traduire, mais qui sont à la fois représentatifs des domaines en question.

3. `http://corpus.byu.edu/coca/`
4. Le corpus vulgarisé a également été utilisé dans nos expériences.

La table 6.4 contient ceux-ci. Le choix de ces termes est également motivé dans la section 6.6.1.

| Corpus | Termes à illustrer |
|---|---|
| Vulcano FR | *basalte, cendre, cratère, cône, débris, dégazage, dôme, fontaine, gaz, jaillir, lave, magma, phase, roche, scorie, téphra, vacuole, volcan, vésicule, éruption* |
| Vulcano EN | *basalt, blobs, cinder, cone, eruption, fountaining, gas, layers, scoria, softball, spongelike, vesicles* |
| Cancer FR | *curage, dépistage, ganglion, carcinome, séquelle, rechute, envahissement, douleur, zonectomie, surdosage, récidive, exérèse, morbidité, guérison, anomalie* |
| Cancer EN | *dissection, screening, node, carcinoma, relapse, involvement, distress, lumpectomy, boost, recurrence, excision, morbidity, recovery, abnormality* |

TABLE 6.4 – *Liste des termes à illustrer dans les deux corpus vulcanologie (Vulcano) et cancer du sein (Cancer)*

## 6.3 Évaluation manuelle des connaissances conceptuelles

Après avoir présenté les ressources utilisées, nous détaillons maintenant l'évaluation manuelle (de point de vue linguistique) des contextes obtenus par les PC. Nous rappelons tout d'abord la fiabilité des PC, puis nous décrivons les consignes de l'évaluation et les résultats obtenus.

### 6.3.1 Fiabilité des PC

Selon Pantel et Pennacchiotti (2008), un patron n'est fiable que lorsqu'il est attesté sur un corpus de grande taille, ayant éventuellement un rappel faible mais surtout une précision élevée. Pour ces auteurs, la fiabilité d'un patron se mesure, dans le corpus, par la qualité des relations entre les termes. Cette fiabilité tend à privilégier des patrons très précis. Les auteurs s'en servent pour juger de la pertinence de patrons génériques à adapter au corpus.

### 6.3.2 Validation manuelle des CRCC

La table 6.5 contient des exemples de contextes fournis par notre méthode mettant en œuvre des PC d'hyperonymie. Le CRCC candidat « *Les **cendres** sont les principaux **produits volcaniques** émis par les volcans explosifs de la ceinture de feu du Pacifique* » explicite une relation valide d'hyperonymie définissant le terme *cendre*. Il s'agit alors d'un CRCC valide. En ce qui concerne le deuxième cas « *L'asymétrie entre les deux **seins** est la **séquelle** la plus souvent rencontrée après traitement conservateur* », la relation d'hyperonymie est invalide. En effet, le terme *séquelle* n'est pas un hyperonyme de *sein*, ainsi nous considérons ce contexte comme non intéressant.

Même s'ils peuvent contenir d'autres relations sémantiques intéressantes, les CRCC candidats sont évalués seulement sur la base du PC (d'hyperonymie ou de méronymie) qui les a repéré.

| Terme à illustrer (X) | Terme du domaine (Y) | Marqueur d'hyperonymie | CRCC candidat |
|---|---|---|---|
| Cendre | produit volcanique | X_ÊTRE_LE_PRINCIPAL_Y | *Les **cendres** sont les principaux **produits volcaniques** émis par les volcans explosifs de la ceinture de feu du Pacifique.* |
| Sein | séquelle | X_ÊTRE_Y_LE_PLUS | *L'asymétrie entre les deux **seins** est la **séquelle** la plus rencontrée après traitement.* |

TABLE 6.5 – *Exemples de CRCC candidats identifiés par des PC d'hyperonymie pour le français*

### 6.3.3 Résultats de validation de CRCC

Les tables 6.6 et 6.7 présentent les résultats obtenus après projection des PC d'hyperonymie et de méronymie sur les corpus de vulcanologie (vulgarisé) et cancer du sein (scientifique). Les termes de la table 6.4 sont rappelés dans la colonne # *Termes à illustrer*. Nous désignons par # *Termes extraits* le nombre de termes intégrant des PC avec au moins un hyperonyme (ou un méronyme), et par # *CRCC candidats* le nombre de CRCC candidats associés aux termes extraits. # *CRCC* indique le nombre de CRCC validés manuellement comme décrit dans la section 6.3.2.

La qualité des CRCC obtenus à partir du corpus vulgarisé de la vulcanologie est satisfaisante, notamment ceux extraits avec des PC d'hyperonymie. Ces résultats sont finalement assez conformes à l'état de l'art (Morin 1999, Malaisé et al. 2004), à savoir un faible rappel des patrons de connaissances au bénéfice d'une bonne précision. Concernant le corpus scientifique cancer du sein, les PC ont gardé le même comportement (précision relativement élevée par rapport au rappel) avec une chute considérable de précision, en particulier pour les PC d'hyperonymie EN qui ne couvrent qu'un seul terme. Cette différence de résultats peut être expliquée par le fait que le discours de vulgarisation se marque par la définition et la description de la terminologie utilisé. En revanche, en discours spécialisé, notamment dans des communications récentes, la terminologie n'est pas toujours introduite par des énoncés définitoires. Seuls les travaux fondateurs, ou introduisant de nouvelles notions, intègrent des énoncés définitoires.

Nous avons dégagé certains enjeux qui peuvent affecter le comportement des PC : le fait d'introduire les termes et leur position dans le PC rend ce dernier rigide et réduit le nombre d'occurrences ; l'un des deux termes peut être remplacé par une anaphore ; les positions des termes sont inversées. Dans l'évaluation, beaucoup de marqueurs peuvent être associés à plusieurs relations comme le signalait Meyer (2001). Par exemple, *X fait partie de Y* peut être interprété comme méronymie ou hyperonymie. La colonne # *Termes extraits* montre que les PC (à l'exception des PC d'hyperonymie) sont présents dans les deux corpus spécialisés même si uniquement 33 à 40 % des termes à illustrer sont retrouvés. Le nombre des CRCC

candidats et celui des CRCC valides montrent une acceptable productivité des PC utilisés dans deux corpus distincts, ainsi qu'une meilleure stabilité en corpus vulgarisé.

| Corpus | # Termes à illustrer | # Termes extraits | # CRCC candidats | # CRCC (P.) |
|---|---|---|---|---|
| **Hyperonymie** | | | | |
| Vulcano FR | 20 | 8 | 21 | 17 (80,95 %) |
| Vulcano EN | 12 | 4 | 14 | 10 (71,42 %) |
| **Meronymie** | | | | |
| Vulcano FR | 20 | 10 | 20 | 13 (65 %) |
| Vulcano EN | 12 | 7 | 25 | 16 (64 %) |

TABLE 6.6 – *Résultats de la projection des PC après une validation manuelle pour le corpus Vulcanologie*

| Corpus | # Termes à illustrer | # Termes extraits | # CRCC candidats | # CRCC (P.) |
|---|---|---|---|---|
| **Hyperonymie** | | | | |
| Cancer FR | 15 | 8 | 14 | 7 (50 %) |
| Cancer EN | 14 | 1 | 2 | 0 ( %) |
| **Meronymie** | | | | |
| Cancer FR | 15 | 5 | 8 | 4 (50 %) |
| Cancer EN | 14 | 3 | 7 | 3 (42,85 %) |

TABLE 6.7 – *Résultats de la projection des PC après une validation manuelle pour le corpus Cancer du sein*

### 6.3.4 Problèmes rencontrés et solutions

Plusieurs problèmes ont été rencontrés lors de la mise en œuvre et l'évaluation des PC. Nous les séparons en deux catégories : liés au pré-traitement et liés à la projection des PC.

#### Pré-traitement

Bien qu'ils soient performants, les outils de segmentation de corpus en occurrences de phrases ne peuvent pas échapper à certaines erreurs. Les parties de textes mal-segmentés peuvent affecter la qualité des contextes extraits. Dans certains cas, les CRCC candidats sont des extraits de phrases incomplètes (mal segmentées), des titres, des extraits de tableaux ou même des légendes et des titres d'images tels que « *Contexte cellulaire et autres anomalies* ». Ces contextes incomplets n'ont pas été validés du fait d'un éventuel manque d'informations et des problèmes de structuration qu'ils peuvent contenir.

#### Projection de PC

La terminologie du domaine a été extraite automatiquement avec TermSuite. En pratique, cet outil propose une liste ordonnée de termes candidats pouvant être ajustée par l'utilisateur. Comme nous n'avons pas appliqué de seuil, certains termes candidats ne le sont pas, et affectent négativement la qualité des PC. Par exemple dans le contexte suivant « *Un exemple*

*classique de ce type est l'éruption »* qui illustre *éruption*, le terme candidat *type* ne peut être considéré comme valide.

Certains PC ne fournissent pas de contextes, ce qui peut être dû à plusieurs raisons :

— **aucune phrase ne contient le PC projeté :** cela signifie que le terme en question n'a pas été illustré dans le corpus par ce PC.

— **le PC est très restrictif :** c'est-à-dire qu'il peut être présent mais dans une forme plus relaxée que celle utilisée au départ. Par exemple, le PC `Y est le X` ne détectera pas « *L'Olympus (ci-contre) est le volcan géant du système solaire* »[5]. Pour cela, les PC qui n'ont pas donné de CRCC candidats ont été « relaxés » en ajoutant la possibilité de contenir d'autres mots (1 à 3 mots, ce nombre est choisi empiriquement). Seulement deux PC ont été relaxés.

— **les termes font partie d'une expression polylexicale :** ceci représente une difficulté lors de la projection et l'évaluation des PC. Deux cas ont été repérés :

1. quand les termes du PC apparaissent au sein d'une de leurs variantes, ils ne seront pas détectés par ce PC. Par exemple, si l'on cherche à illustrer le terme *basalte*, le contexte « *Les basaltes sombres sont les roches volcaniques les plus fréquentes* » ne sera pas pris en compte du fait que *basalte* occurre au sein de sa variante *basaltes sombres*. Pourtant, il s'agit d'un CRCC intéressant. Un autre exemple dans lequel le deuxième terme du domaine n'est pas identifié, est le PC `Y est une X` qui couvrira *L'Italie est une terre de volcans*[6] tandis que le contexte *L'Italie est une grande terre de volcans* ne sera pas identifié. Lorsqu'un PC ne fournit pas de contextes, sa relaxation pourrait résoudre ce problème.
2. quand le terme à illustrer apparaît dans un terme complexe de type NPN, le jugement du contexte devient difficile. Par exemple, dans le contexte « *La Soufrière est un volcan d'arc* ».

## 6.4 Évaluation manuelle des connaissances linguistiques

L'idée est d'utiliser des collocations comme des connaissances linguistiques illustrant un usage typique d'un terme donné en domaine de spécialité. En outre, les contextes procurés par les collocations, peuvent également contenir des relations conceptuelles. Nous étudions la validité de l'hypothèse selon laquelle les collocations procurent une connaissance linguistique dans leurs contextes d'apparition. Nous étudions également la présence des connaissances conceptuelles dans ces contextes. Nous avons réalisé une évaluation manuelle pour laquelle nous présentons les consignes fournies aux annotateurs, ainsi que les résultats obtenus.

5. Avec X = *Olympus*, et Y = *volcan géant*.
6. Avec X = *Italie*, et Y = *terre de volcan*

### 6.4.1 Consignes aux annotateurs

Trois linguistes du CLLE-ERSS partenaires du projet CRISTAL ont annoté les contextes que nous leurs avons fournis. Lorsqu'il y a beaucoup de contextes, ils n'en ont annoté qu'une partie. Les annotateurs se sont mis d'accord sur la compréhension de la tâche à réaliser :

1. vérifier tout d'abord s'il s'agit d'une collocation valide ou non en dehors des contextes proposés. C'est-à-dire, vérifier si l'on peut construire une collocation à partir du terme et son collocatif proposé (éventuellement dans un domaine général). Par exemple, à partir du terme *fibre*, et son collocatif proposé *riche*, l'association *riche en fibre* peut être considérée comme une collocation valide (ou acceptable). Aucune contrainte n'est fixée, la consultation de ressources externes (Web, corpus...) est libre.
2. si la collocation est valide, la démarche est de vérifier, par rapport à la compréhension du terme visé, s'il s'agit tout d'abord d'un contexte conceptuel, sinon linguistique ou le cas échéant non intéressant (*cf*. section 6.4.2).
3. si la collocation n'est pas valide, le contexte ne sera pas validé.

Dans plusieurs collocations proposées par le z-score, telles que *cholestérol* et *mis*, l'association *cholestérol mis* ne forme pas une collocation valide. Une vérification manuelle permet d'affiner la qualité des collocations identifiées. En pratique, nous retiendrons les collocations sans évaluation. Cependant, dans ce travail, nous commençons par évaluer les collocations identifiées pour chaque terme de manière à réaliser une évaluation fine de notre méthode.

### 6.4.2 Validation manuelle des CRCL

Les contextes candidats ont été annotés par des linguistes, et validés par rapport aux collocations qui ont permis de les détecter. Nous désignons par CRCL valide, un contexte qui contient une collocation (terme à illustrer, collocatif) valide. Dans le cas contraire, le contexte est considéré comme non intéressant (ou non riche). En plus des collocations, les CRCL valides peuvent également contenir des relations sémantiques intéressantes. Nous distinguons alors deux types de CRCL valides : CRCL « pure »[7] ; et CRCL contenant aussi des CRCC.

— **Un CRCL** est un contexte contenant seulement le terme visé $T$ et son collocatif (dans une collocations valide dans ce contexte), à l'exception d'autres termes du domaine auxquels $T$ est conceptuellement lié. Un CRCL doit être grammaticalement bien formé, c'est-à-dire bien structuré et compréhensible.

— **Un CRCC :** est un contexte contenant le terme visé $T$ et son collocatif (la collocation doit être valide dans ce contexte), ainsi que d'autres termes du domaine étudié qui sont conceptuellement liés avec $T$.

7. Qui ne contiennent pas d'informations conceptuelles.

Nous désignons par termes conceptuellement liés, tous termes reliés par un PC ou manifestant une relation conceptuelle entre eux. Dans la section 6.3, nous nous sommes limités aux relations d'hyperonymie et de méronymie. Ici, nous prenons en considération toute relation sémantique qui peut apparaître dans ce contexte et qui peut aider à comprendre le terme en question, sans pour autant définir la nature ou le type de cette relation.

Les exemples de la table 6.8 sont obtenus en appliquant la méthode basée sur les collocations. Les exemples **i)** et **iii)** contiennent des collocations valides, et illustrent respectivement des relations de causalité entre *glycémie* et *insuline* et d'hyponymie entre *glycémie* et *bon cholestérol* et *mauvais cholestérol*. Ils correspondent alors à des CRCC.

Contrairement à **i)**, les exemples **ii)** et **iv)** sont seulement des CRCL (car les collocations sont valides) même s'ils contiennent d'autres termes (*diabète de type 1*, *dépistage*...) du domaine n'étant pas liés conceptuellement avec *glycémie*. Ici la validation des relations conceptuelles dépend seulement de l'interprétation de l'annotateur.

— **Un non CRC :** est CRCL candidat fourni par une collocation non valide (dans ce contexte).

Cette évaluation conduit alors à considérer les contextes valides fournis par les collocations comme des CRC à minima CRCL, et au mieux CRCL et CRCC.

| Collocation | CRC candidat | CRCL | CRCC |
|---|---|---|---|
| (**Glycémie**, élever) | **i)** *Il y a une régulation un peu à la manière d'un thermostat : si la **glycémie** s'**élève**, le pancréas fabrique davantage d'insuline pour permettre l'utilisation du sucre.* | | x |
| | **ii)** *Le dépistage le diabète de type 1 ne pose pas de problème de dépistage car la **glycémie** s'**élève** de façon brutale et il y a des symptômes évidents.* | x | |
| (**Cholestérol**, mauvais) | **iii)** *Il existe deux sortes de **cholestérol** : le bon **cholestérol** et le **mauvais cholestérol**.* | | x |
| | **iv)** *Il est difficile d'ignorer qu il y a du **mauvais cholestérol**.* | x | |

TABLE 6.8 – *Exemple d'évaluation de CRCL candidats (les termes visés sont en gras. Les termes du domaine sont soulignés)*

.

### 6.4.3 Résultats des collocations

La méthode d'identification des CRCL a été appliquée sur les corpus vulcanologie (VU) et cancer du sein (SC).

Les tables 6.9 et 6.10 présentent les résultats obtenus pour les collocations de type (terme, adjectif), (terme, nom) et (terme, verbe) pour les corpus vulcanologie et cancer du sein. # *Termes extraits* est le nombre de termes couverts par les collocations. *CRCL* représente, les pourcentages des CRCL valides. *CRCC* contient les pourcentages des CRCC (selon les consignes 6.4.2) parmi les CRCL. *Non CRC* est le pourcentage de contextes non-intéressants. Les décisions prises résultent d'une discussion entre les trois annotateurs. Selon les retours de ces derniers, le jugement sur la collocation « hors contexte » a été assez facile à émetre, celui sur le contexte est beaucoup plus aléatoire. Si le contexte est à la fois intéressant conceptuellement et linguistiquement, la catégorie « conceptuel » a été privilégiée. Lorsque les contextes fournis par les collocations sont très nombreux (181 pour *(lave, coulée)*), seulement les dix premiers ont été annotés. Les contextes marqués non CRC (non riches) peuvent relever de trois cas de figure :

1. cas le plus courant : la collocation y figure, mais il n'y a pas d'information supplémentaire intéressante (étant donné qu'il contient au moins la collocation, les annotateurs pourrait le considérer par défaut comme intéressant linguistiquement, mais ils n'ont pas raisonné ainsi) ;
2. cas très rare : la collocation n'y figure pas (c'est un « vrai » mauvais contexte). Ex : pour (*lave, visqueux*), le contexte « *Un dôme de lave visqueux...* » (*visqueux* modifie *dôme* plutôt que *lave*) ;
3. le contexte est mal structuré, difficilement compréhensible (problème d'anaphore, phrase interrompue...).

Nous focalisons notre analyse des résultats sur le type des connaissances illustrées ainsi que sur la cohérence entre les résultats des corpus étudiés. Dans le corpus vulcanologie, nous nous sommes limités à l'évaluation des contextes fournis par les 5 premières collocations selon le Z-score. Ce choix nous a permis de focaliser notre étude sur les collocations les plus pertinentes (selon le Z-score), et de réduire la complexité de l'exercice qui est très coûteux en termes de temps. Pour le corpus cancer du sein, nous avons supprimé cette contrainte et évalué toutes les collocations obtenues par le Z-score. Ceci explique la différence importante de la colonne # *CRCL candidats* des deux tables 6.9 et 6.10.

Les résultats montrent que les contextes obtenus grâce aux collocations sont majoritairement des CRCL valides. Nous expliquons la meilleure précision dans le corpus cancer du sein, par le fait que le discours scientifique représente un réservoir de contextes illustrant des usages « typiques » et une terminologie plus technique. En outre, les CRCL contiennent une quantité considérable de connaissances conceptuelles, notamment dans le corpus cancer du sein. Ceci montre que les relations sémantiques sont plus fréquentes dans les voisinages (*i.e* contextes) des collocations en corpus scientifique qu'en corpus vulgarisé. Les résultats dans les deux corpus

ainsi que dans les deux langues sont cohérents. En effet, les collocations se comportent pareillement selon les catégories grammaticales.

Nous tenons à mentionner que certains CRCC fournis par les collocations contiennent des relations d'hyperonymie ou de méronymie mais qui n'ont pas été extraites par les PC utilisés auparavant. Dans la plus part des cas, les termes en relation sont éloignés dans les contextes. Dans d'autres CRCC, le lien n'était pas explicite, d'où la validation s'est appuyée seulement sur l'interprétation de la relation. Certains contextes ont été validés car ils contiennent des liens que l'annotateur a jugés utiles, sans pour autant décrire le type de ce lien ou de la relation.

| **Vulcan.** | **# Termes à illustrer** | **# Termes extraits** | **# CRCL candidats** | **CRCL** | **(CRCC)** | **Non CRC** |
|---|---|---|---|---|---|---|
| | | | **ADJ** | | | |
| Français | 20 | 18 | 74 | 58,11 % | (45,95 %) | 41,81 % |
| Anglais | 12 | 10 | 41 | 56,09 % | (41,46 %) | 43,90 % |
| | | | **NOM** | | | |
| Français | 20 | 18 | 83 | 59,04 % | (31,33 %) | 40,96 % |
| Anglais | 12 | 10 | 43 | 72,09 % | (41,86 %) | 27,91 % |
| | | | **VER** | | | |
| Français | 20 | 11 | 18 | 27,77 % | (60 %) | 72,23 % |
| Anglais | 12 | 9 | 14 | 42.85 % | (66,66 %) | 57.15 % |

Table 6.9 – *Évaluation des contextes extraits par les collocations pour la Vulcanologie*

| **Cancer.** | **# Termes à illustrer** | **# Termes extraits** | **# CRCL candidats** | **CRCL** | **(CRCC)** | **Non CRC** |
|---|---|---|---|---|---|---|
| | | | **ADJ** | | | |
| Français | 15 | 14 | 1 151 | 100 % | (56,93 %) | 0 % |
| Anglais | 14 | 14 | 415 | 89,73 % | (71,79 %) | 10,25 % |
| | | | **NOM** | | | |
| Français | 15 | 13 | 292 | 86,9 % | (51,19 %) | 13,09 % |
| Anglais | 14 | 14 | 494 | 84,43 % | (57,77 %) | 15,55 % |
| | | | **VER** | | | |
| Français | 15 | 10 | 158 | 60,97 % | (46,34 %) | 39,04 % |
| Anglais | 14 | 13 | 56 | 59,25 % | (44,44 %) | 40,74 % |

Table 6.10 – *Évaluation des contextes extraits par les collocations pour le Cancer du sein*

## 6.5 Synthèse : stratégie unifiée

Si nous combinons maintenant les résultats obtenus, nous pouvons constater à la lecture des tables 6.11 et 6.12 que nous sommes en mesure de proposer des CRC pour l'ensemble des termes de la liste de référence. La qualité de ces CRC est autour d'environ 70 %, malgré la complexité de l'exercice. Les résultats obtenus permettent de mettre en évidence l'utilité de l'exploitation des collocations pour résoudre le problème du faible rappel des PC.

Il serait intéressant d'ordonner les contextes obtenus afin de n'en conserver que les plus intéressants. Dans un premier temps, il serait privilégié de proposer en premier lieu les CRC issus des patrons de connaissances puis dans un second temps ceux issus des collocations.

| Corpus | # Termes à illustrer | # Termes extraits | # CRC candidats (sans doublons) | # CRC valides (P.) |
|---|---|---|---|---|
| Français | 20 | 20 | 143 | 100 (69,93 %) |
| Anglais | 12 | 10 | 97 | 67 (69,07 %) |

TABLE 6.11 – *Tableau récapitulatif de la combinaison des deux méthodes pour le corpus Vulcanologie*

| Corpus | # Termes à illustrer | # Termes extraits | # CRC candidats (sans doublons) | # CRC valides (P.) |
|---|---|---|---|---|
| Français | 15 | 15 | 1 659 | (66,28 %) |
| Anglais | 14 | 14 | 1 008 | (65,43 %) |

TABLE 6.12 – *Tableau récapitulatif de la combinaison des deux méthodes pour le corpus Cancer du sein*

## 6.6 Évaluation expérimentale en traduction

Dans le cadre du projet CRISTAL, cinq expérimentations ont été réalisées. Certaines ont servi à la vérification des conditions expérimentales et à l'ajustement des protocoles adoptés (Planas et al. 2014, Josselin-Leray et al. 2014). Nous présentons celles qui ont permis de valider les résultats et tirer les conclusions du projet.

Les principaux objectifs des premières expérimentations (mars 2014) qui ont précédées nos évaluations étaient (i) de vérifier l'utilité des CRC intéressants du point de vue linguistique dans un exercice de traduction ; et (ii) d'identifier les types de connaissances (CRCC ou CRCL) qui focalisent plus l'attention des traducteurs. Ces expériences ont fait l'objet des travaux de Josselin-Leray et al. (2014) qui montrent l'intérêt d'extraire des contextes riches, et valident l'hypothèse qui a donné lieu au travail de cette thèse. Nous considérons cette expérience comme une preuve de concept pour la suite de nos travaux.

Les deuxièmes expériences (avril 2015) avaient pour but (i) d'évaluer l'interface Libellex développée dans le cadre de CRISTAL, et (ii) d'étudier l'utilité des CRC fournis par nos méthodes par rapport à l'aide à la traduction terminologique. Nous présentons tout d'abord les données utilisées, ensuite nous détaillerons les expériences réalisées.

### 6.6.1 Données expérimentales

Les expériences consistent à traduire, de l'anglais vers le français, un texte de vulgarisation scientifique sur la volcanologie et contenant environ 150 mots (table 6.13). Le choix de ce texte a été basé sur sa bonne structuration (les deux phases de la construction d'un *cinder cone* sont décrites) ainsi que les termes qu'il contient. La traduction de certains de ces termes pourrait être complexe pour un traducteur, même s'ils ne sont pas très spécialisés (*basalt cinder cone, fountaining stage*...). Le texte contient également des collocations qui sont particulièrement difficiles à traduire comme *bubble off* (en FR *éclater*).

Les termes sur lesquels portent ces expériences sont des termes sur lesquels la définition des CRC pouvait être appliquée. Par exemple, il serait

inutile de proposer des CRCC à des termes familiers qui ne sont pas spécifiques au domaine de la volcanologie. Les termes en question sont : *basalt magma, blobs, cinder, cinder cone, fountaining, scoria, vesicles* ainsi que leurs équivalents possibles.

| |
|---|
| *CINDER CONES*<br>*Basalt cinder cones are the most common kind of volcano worldwide. They are also some of the smallest volcanoes. A typical eruption goes through two stages. The first is called the fountaining stage. When the basalt magma first breaks out at the surface, the dissolved gases bubble off vigorously enough to carry blobs of magma into the air with them. The blobs may rise up 2,000 feet or more.*<br>*During their flight through the air, the blobs cool to solid pieces called cinders (also called scoria), which are typically no bigger than a softball. These spongelike rocks contain holes, called vesicles, where gas bubbles got trapped inside. As the cinders fall back to Earth, they form layers that pile up into a cone-shaped hill. The very top, where the magma spewed out, usually ends up with a small crater. Once most of the gassy magma has erupted, fountaining stops.* |

TABLE 6.13 – *Le texte source à traduire*

### 6.6.2 Expérimentations préalables

#### Protocole d'expérimentations

42 étudiants en deuxième année du Master de traduction à l'Université Catholique de l'Ouest ont participé aux expériences. Les contextes ont soigneusement été choisis au préalable par une équipe de linguistes avant d'être proposés aux traducteurs dans un ordre aléatoire en langue source (et en langue cible) pour chaque terme (et ses traductions possibles). Le choix de ces contextes a été basé sur une analyse de concordances des corpus, complétée par des recherches sur des sites de volcanologie sur internet. Des contextes qui ne sont pas riches en connaissances ont été également proposés. Au cours du processus de traduction, les participants devaient choisir les contextes qui étaient les plus utiles pour eux lors de la traduction. Le déroulement des expériences a été enregistré grâce à des enregistrements vidéos des écrans des traducteurs ainsi que les traces de leurs claviers.

Les traducteurs ont été répartis sur deux groupes dont un seul était peu familier avec le domaine de la volcanologie. Les expériences ont duré 2 heures pour traduire le texte, choisir les contextes pertinents et remplir un questionnaire en ligne sur les principales difficultés de la traduction et de l'utilité de leurs ressources habituelles (dictionnaires et banques terminologiques...) et des CRC.

#### Observations et résultats

Pour la langue source, 48 % des contextes proposés ont été choisis par au moins un participant (108 contextes) contre 36 % pour la langue cible (152 contextes). Ceci montre que les contextes sont perçus comme utiles, mais les données sont très dispersées, puisque environ 40 % des contextes sélectionnés ont été choisis par un seul participant dans les deux langues. Afin d'affiner l'interprétation des résultats, l'accent a été mis sur les 20 contextes les plus sélectionnés dans les deux langues : chaque contexte a été choisi entre 4 et 14 fois.

| | **CRC** | **CRCC (définitions)** | **CRCL** |
|---|---|---|---|
| **Langue source** | | | |
| Tous les contextes | 69 % | 25 % (11 %) | 52 % |
| 20 contextes plus choiss | 90 % | 65 % (60 %) | 25 % |
| **Langue cible** | | | |
| Tous les contextes | 70 % | 29 % (14 %) | 49,5 % |
| 20 contextes plus choiss | 90 % | 55 % (40 %) | 35 % |

TABLE 6.14 – *Exploitation des CRC en traduction*

La table 6.14 établit une comparaison entre ce sous-ensemble de contextes et tous les contextes qui ont été mis à la disposition des expérimentations. Cette table montre tout d'abord que la majorité des contextes considérés comme utiles par les participants sont des contextes riches en connaissances (90 %). Ensuite, les participants montrent une forte préférence pour les contextes CRCC, principalement les définitions, plutôt qu'aux CRCL. Cependant, en analysant les deux langues, nous pouvons constaté que la répartition entre CRCL et CRCC est différente lorsque les utilisateurs explorent les CRC sources (en anglais) et cibles (français). Les traducteurs semblent privilégier les CRCC et les définitions en langue source. Cette observation valide l'hypothèse de base selon laquelle les CRC utiles du point de vue linguistique sont également utiles pour l'aide à la compréhension et à la traduction terminologique. Ces résultats sont encourageants d'autant plus qu'ils suggèrent différentes façons dont le traducteur utilise ces contextes en langues source et cible : les CRCL devraient être plus utiles lors de la vérification de l'usage des termes en langue cible.

### 6.6.3 Expérimentations finales

#### Protocole d'expérimentations

Cette expérience a eu lieu en avril 2015, dans le but d'évaluer l'interface destinée aux traducteurs et développée dans le cadre du projet CRISTAL. L'expérience a été réalisée dans une salle informatique dédiée, avec la participation de huit traducteurs professionnels ayant chacun plus de 10 ans d'expérience, de toutes spécialités confondues.

Les traducteurs devaient traduire le texte de la table 6.13, avec des conditions ergonomiques plus favorables que celles des expériences préalables. L'interface Libellex intègre les contextes proposés, ainsi que des dictionnaires [8] en ligne et un dictionnaire numérisé selon le format $A = B$, où $B$ est une définition. Contrairement aux expérimentations préalables, tous les contextes ont été fournis automatiquement par nos méthodes : des CRC candidats. Pour certains termes, le nombre de contextes proposés est très important. Un compromis a été d'en proposer 10 au maximum : idéalement, 5 conceptuels et 5 linguistiques. Pour cela les contextes ont été pré-traités selon la stratégie suivante :

1. ajouter automatiquement un score aux contextes proposés : il s'agit du nombre de termes à traduire présents dans le CRC candidat ;
2. proposer les 5 premiers contextes (d'abord conceptuels, ensuite linguistiques) selon ce score ;
3. si le nombre de contextes est inférieur à 5, compléter aléatoirement avec l'autre type de CRC (linguistique ou conceptuel) sans prendre en considération le score ;
4. si le nombre de contextes est encore inférieur à 5, ne rien faire.

**Observations et résultats**

À la fin de ces expériences, les traductions et les activités des traducteurs ont été enregistrées. Les questionnaires ont été remplis par chacun des participants, deux entretiens en groupe (deux groupes de quatre) et un entretien individuel général supplémentaire ont été réalisés. Le traitement des données récoltées nécessite un post-traitement comprenant la compilation de chaque *log* de traducteur pour déterminer pour chaque terme recherché, quelles ressources ont été sollicitées, quels CRC et quel est leur type. Cette tâche nécessite également une compilation inter-traducteurs pour comparer et détecter les tendances.

Ce traitement est en cours. Nous nous limitons alors aux retours des traducteurs pour tirer des premières conclusions de cette expérience. Il a été mentionné que certains CRC sont plus utiles que d'autres, selon les phases de l'acte de traduction. Par exemple, les définitions (des CRC conceptuels avec des PC d'hyperonymie) pour la phase de compréhension du texte, ou les collocations dans la phase de production. Les contextes qui contiennent plusieurs connaissances (PC, collocations et d'autres termes) peuvent être qualifiés de « très riches » du fait qu'ils peuvent être utilisés à différents stades de traduction. La plupart des contextes ont été utiles en complément des autres ressources sollicitées.

8. Ressources en ligne : synonymes, Linguee, R&C, le Grand Dictionnaire Terminologique et Termium.

## 6.7 Conclusion

Dans ce travail, nous avons proposé de mettre en œuvre la notion de contextes riches en connaissances afin d'extraire directement de corpus spécialisés des exemples illustrant le fonctionnement des termes. Ces CRC, s'appuyant sur des patrons de connaissances et des collocations, permettent d'accéder à des connaissances linguistiques et conceptuelles. L'originalité de notre approche est de considérer l'ensemble des CRC disponibles à la différence des travaux existants se restreignant soit à des patrons de connaissances pour extraire des définitions (Marshman 2014) soit à des collocations pour extraire des exemples (Kilgarriff et al. 2008). Une première évaluation linguistique, a permis d'identifier des connaissances conceptuelles utiles, en plus des connaissances linguistiques fournies principalement par les collocations. Les résultats préliminaires également issus des expériences ont confirmé nos hypothèses qui postulent que les contextes riches du point de vue des linguistes peuvent être perçus comme riches par les traducteurs en complément de ses ressources habituelles. Il serait intéressant de pouvoir réduire le nombre proposé de CRC en cherchant à proposer systématiquement pour un terme à illustrer un exemple de CRC linguistique et un autre CRC conceptuel. Pour ce faire, il sera nécessaire de pouvoir associer à chacun de ces CRC un score selon la densité terminologique du contexte ou un score de confiance pouvant être par exemple en fonction du patron de connaissances déclenché dans un cas et de l'ordonnancement global des collocations dans l'autre.

# Troisième partie

# CRC bilingues

# Extraction de CRC bilingues

7

Sommaire

## 7.1 Introduction

Tout travail de traduction, notamment celui de traduction spécialisée, nécessite la mobilisation d'informations terminologiques accessibles à travers des glossaires et des outils de gestion dédiés. Ce travail met en œuvre des compétences linguistiques tant en termes de rédaction que de vérification qui peuvent parfois s'exercer comme un métier à part entière. Enfin, un certain nombre de traducteurs endossent également une casquette de terminologues, de chefs de projets ou encore de réviseurs.

Dans ses travaux, Gouadec (2002) regroupe les activités de traduction en trois grandes phases : pré-traduction, traduction et post-traduction. La phase de **pré-traduction** consiste à préparer les fichiers et les ressources nécessaires à la traduction. La phase de **traduction** est menée par des traducteurs ou des rédacteurs spécifiquement formés pour traduire des documents très spécialisés tels que des notices ou des manuels d'utilisation. La **post-traduction**, selon Gondoin (2007), comprend principalement la réintégration des chaînes traduites dans le format d'origine, ainsi qu'un contrôle de l'intégrité des documents ainsi construits.

Au-delà de la traduction d'un texte par un professionnel qualifié, il devient de plus en plus courant d'effectuer un contrôle de la qualité du texte traduit. Chacun reconnaîtra qu'un traducteur, qu'il soit indépendant ou salarié, ne peut fournir en permanence des traductions irréprochables. Le manque de contextes lors de la phase de traduction renforce la nécessité de contrôles tels que la révision, qui permet d'assurer la cohérence globale du document produit.

De manière officielle, la révision a été révélée grâce aux normes German DIN 2345, European EN 15038 et ISO 17000 prévoyant l'obligation de faire réviser toute traduction. Selon Robert (2012, p. 95), l'exercice de révision comporte deux axes principaux : la révision bilingue dans laquelle le réviseur compare le texte original avec le texte traduit ; et la révision monolingue où seul le texte traduit est révisé. Ces deux révisions peuvent être effectuées par le traducteur lui-même, afin d'améliorer les traductions qu'il produit, ou comme le recommandent les normes européennes, par un autre traducteur aussi appelé réviseur.

Le choix de la méthode de révision est le plus souvent un compromis entre la complexité et la difficulté de la traduction d'un coté, et les ressources disponibles (temps et personnel) de l'autre coté. Néanmoins, le réviseur ne s'en tient pas uniquement à la révision monolingue : le document original sert de support en cas de doute.

La révision est également connue sous d'autres appellations (Robert, Isabelle 2008, p. 4) telles que relecture, vérification, amélioration ou aussi QC-ing (*Quality Control*). Ici, nous retenons le terme révision et nous nous penchons plus particulièrement sur la révision bilingue dans laquelle le réviseur doit confirmer ou infirmer la traduction attestée par le traducteur (Delisle et al. 1999, p. 71) pour une éventuelle correction.

## 7.2 Révision en domaine de spécialité

Dans une étude effectuée par Morin-Hernandez (2009, p. 143), 90 % des traducteurs français ont répondu par un indice de 1 à 3 (sur une échelle de 5 commençant par 1) qu'une erreur terminologique a un impact majeur sur le travail de traduction. En effet, la terminologie est un facteur important qui agit directement sur la qualité du document produit, en particulier lorsqu'il s'agit de traduction en domaine de spécialité. La table 7.1 contient deux exemples qui illustrent concrétement l'embûche rencontrée lors d'un exercice de révision en domaine de spécialité. Considérons la traduction des termes anglais *blob* et *cinder* dans les textes suivants à traduire :

| **Texte à traduire (EN)** | **Traduction produite (FR)** |
|---|---|
| a) *When the basalt magma first breaks out at the surface, the dissolved gases bubble off vigorously enough to carry* ***blobs*** *of magma into the air with them. The* ***blobs*** *may rise up 2,000 feet or more.* | *Lorsque le magma basaltique explose à la surface, les gaz dissous partent de manière si puissante qu'ils emportent avec eux des* ***gouttes*** *de magma. Les* ***gouttes*** *peuvent être expulsés à au moins 600 mètres de hauteur.* |
| b) ***CINDER*** *CONES*<br>*Basalt* ***cinder*** *cones are the most common kind of volcano worldwide.* | *CONES DE* ***SCORIES***<br>*Les cônes de* ***débris*** *basaltiques sont un des types de volcans les plus communs au monde.* |

Table 7.1 – *Exemples de traductions proposées contenant les termes blob et cinder*

Nous constatons à partir du cas a) de la table 7.1 que la traduction du terme *blob* dans le texte à traduire en français n'est pas évidente. Bien qu'en langue générale la traduction commune de *blob* soit *goutte*, la traduction la plus appropriée en vulcanologie est *projection*. Concernant b), la traduction du terme *cinder* (EN) dans un contexte de vulcanologie peut s'avérer problématique pour le traducteur. En effet, deux traductions sont possibles pour le traducteur : *scorie* et *débris*. Une autre traduction *cendre* pourrait être proposée. La traduction en langue générale de *cinder* est bien *cendre*, qui est une ancienne appellation de *scorie*. Toutefois, *cendre* peut également correspondre à la traduction de *ash* (EN). Ici, la phase de révision (voire d'une auto-révision) est indispensable pour prendre la meilleure décision qui est *scorie*.

Dans ces deux cas précédents, l'accès à des contextes contenant des voisinages typiques ou renseignant sur les relations conceptuelles entre les termes en question (*blob* en anglais, ainsi que *goutte* et *projection* en français) et les autres termes du domaine, est essentiel pour le réviseur. Nous parlons des contextes riches en connaissances. De son coté, le réviseur dispose d'un certain nombre de moyens permettant de garantir, autant que faire se peut, la qualité de traduction à laquelle il s'engage. Ces moyens sont les outils et les ressources mises à sa disposition, tels que les dictionnaires, les concordanciers et les corpus de spécialité (comparables et parallèles).

Dans ce chapitre, nous nous intéressons à la problématique d'aider à la révision en traduction spécialisée (de l'anglais vers le français) dans laquelle les traductions ont initialement été attestées par le traducteur dans une étape de production. Dans un premier temps, nous présentons un prototype de concordancier bilingue qui permet de saisir un terme et sa traduction, et fournit des contextes riches en connaissances à partir de corpus comparables spécialisés. Dans un second temps, nous adoptons une approche contrastive d'expérimentation : partant d'une situation de base, dans laquelle le traducteur dispose d'un texte source et sa traduction produite en amont, nous allons observer si la mise à disposition de ce nouveau prototype de concordanciers bilingues en corpus comparables spécialisés, aiderait à la correction des erreurs en révision et ainsi d'améliorer la qualité de la traduction.

## 7.3 Concordanciers bilingues

Il n'est pas rare qu'un même client sollicite le même service de traduction pour traduire à des moments différents des documents distincts, mais qui renferment la même terminologie, voire plusieurs passages communs. Pour cette raison, la plupart des services de traduction archivent plus ou moins systématiquement les documents qu'ils produisent. Face à un nouveau document à traduire, un traducteur disposant des textes apparentés, avec leurs traductions, pourra réutiliser ceux-ci à l'aide d'outils dédiés pour récupérer les traductions réciproques. Cette idée nous mène plus loin que la simple récupération de phrases traduites. Dans ce cas, l'outil qui permet d'observer les (paires de) termes et expressions et leurs traductions dans leurs contextes « bilingues », serait une ressource précieuse pour les traducteurs (Bowker et Barlow 2004) : il s'agit des concordanciers bilingues.

### 7.3.1 Intérêt des concordanciers bilingues

Avant d'aborder le sujet des concordanciers bilingues, il est nécessaire de présenter les concordances monolingues (appelés également unilingues). Traditionnellement, un concordancier monolingue liste toutes les occurrences d'un mot donné en entrée. Chaque occurrence apparaît dans un segment de texte, typiquement au milieu. Celui-ci représente un « contexte » de l'occurrence, d'où le synonyme anglais *keywords in context*, ou KWIC. L'utilisation d'un concordancier devient particulièrement intéressante à partir du moment où la recherche manuelle d'occurrences de mots dans les corpus devient laborieuse. En effet, plus ce terme est fréquent, plus le nombre de contextes devient difficile à manier et à consulter. Dans ce cas, le concordancier apporte une solution à ces problèmes, et permet un affichage et une sélection plus flexibles des contextes.

Les concordanciers bilingues sont similaires aux concordanciers monolingues. Ils permettent d'afficher, pour un terme (ou un mot) saisi, une

paire de contextes qui sont des traductions réciproques. La mise en place d'un concordancier bilingue est plus complexe que celle d'un concordancier monolingue, d'autant plus qu'un problème de « contexte bilingue » se pose. Le concordancier monolingue affiche soit un nombre fixe de caractères ou de mots au voisinage de l'occurrence du terme (ou du mot) en question, soit un contexte plus structuré, tel que la phrase ou le paragraphe contenant cette occurrence. Si l'on souhaite afficher en supplément la traduction de ce contexte, il faut être en mesure de la localiser exactement, ce qui n'est pas toujours possible pour une partie quelconque du texte. Ensuite, le laps de temps minimum requis pour localiser la traduction d'un contexte donné, même de façon approximative, est proportionnel à la taille des deux textes, de telle sorte qu'il n'est pas possible d'effectuer cette opération en temps réel. Il faut donc pré-calculer les correspondances et conserver cette information avec le texte, d'une manière qui assure des accès rapides.

### 7.3.2 Exemple de concordanciers

La plupart des concordanciers bilingues sont plutôt multilingues du fait qu'ils ne sont pas liés à des langues en particulier. Ils sont conçus pour fonctionner avec une ou plusieurs paires de langues. Ces outils permettent aux traducteurs en s'appuyant sur des corpus parallèles de « trouver des informations qui pourraient les aider à compléter une nouvelle traduction » (Bowker et Barlow 2004). Parmi les concordanciers bilingues les plus connus sur le marché, nous citons :

— ParaConc (Barlow 2002) : sa méthode d'alignement prend la phrase comme unité de base, et s'appuie sur l'algorithme Gale-Church (Gale et Church 1993). ParaConc offre une fonction de recherche pour l'analyse de textes parallèles, ce qui permet d'observer les résultats de la recherche dans une mise en page KWIC dans les deux langues, pour trier les contextes et de passer d'une langue à l'autre.

— TransSearch (Bourdaillet et al. 2010) : est un service basé sur le Web qui prend également la phrase comme unité d'alignement. Il donne accès à une base de données de traduction contenant des millions de phrases traduites en anglais, en français et en espagnol. TranSearch affiche les résultats en deux colonnes : une colonne contenant des textes sources dans lesquels le mot ou l'expression apparaît, et une seconde colonne contenant la traduction de ces phrases. Il s'agit d'un moteur de recherche qui pourrait être comparé à Google, mais appliqué à des textes sources avec leurs traductions.

— Find Bi-Text Advantage : Beetext Find a mis sur le marché trois versions de ce concordancier : pour les *freelancers*, pour les traducteurs employés par de petites entreprises, et pour les traducteurs employés par de grandes entreprises. Cet outil permet à plusieurs utilisateurs de collaborer en même temps et d'afficher les documents dans leur mise en page d'origine en complément des contextes.

### 7.3.3 Fonctionnement de concordanciers bilingues

Nous décrivons le fonctionnement des concordanciers bilingues en référence à ParaConc qui est un exemple représentatif de cette catégorie d'outils. Pour pouvoir utiliser ParaConc, les textes sources et cibles doivent être alignés en amont. Un processus d'alignement semi-automatique est inclus dans le concordancier pour préparer des textes qui n'ont pas été préalablement alignés. La première partie de ce processus d'alignement se déroule en trois étapes : les textes sont tout d'abord alignés au niveau des titres, puis au niveau du paragraphe, et enfin au niveau de la phrase. Le logiciel se base également sur la structuration des fichiers pour aligner les paragraphes. L'alignement des phrases est effectué grâce à l'algorithme Gale-Church (Gale et Church 1993). Afin d'ajuster l'alignement, l'utilisateur a la possibilité d'examiner les segments alignés et de fusionner ou découper des segments particuliers, selon ses besoins. Les unités alignées restent affichées dans le texte qui les englobe.

Une fois que les textes sont alignés, le traducteur peut consulter le corpus et récupérer tous les exemples (*i.e* contextes) d'un mot ou d'une expression à partir du corpus. Les lignes de concordance peuvent être triées de différentes manières (les contextes cibles ou sources d'abord) afin de regrouper les phrases similaires ensemble et faciliter le repérage de modèles linguistiques. En cliquant sur une ligne de concordance, cela mettra en évidence cette ligne ainsi que le segment de texte correspondant. Il est également possible d'utiliser une fonctionnalité qui présente une liste de traductions candidates dans la fenêtre des résultats en langue cible. Ces traductions candidates peuvent être sélectionnées, elles seront ensuite mises en évidence dans les résultats. Enfin, des recherches plus avancées peuvent également être effectuées si nécessaire telles que : recherche d'expressions régulières, selon la partie du discours, etc.

Bowker et Barlow (2004) affirment qu'il serait particulièrement intéressant pour les traducteurs de pouvoir effectuer une recherche parallèle, c'est-à-dire en leur permettant d'entrer à la fois un mot (ou un terme) en langue source, et son équivalent en langue cible, pour récupérer uniquement les occurrences qui correspondent à ces deux entrées.

### 7.3.4 Limites des concordanciers bilingues

Selon Bowker et Barlow (2004), certaines limites sont souvent associées aux concordanciers bilingues : i) le statut linguistique de l'élément de la recherche ; et ii) le processus d'alignement.

Les concordanciers bilingues sont généralement conçus pour rechercher des mots ou des expressions très courtes. Les limites concernant les mots simples portent sur le fait que certains concordanciers prennent en compte différentes formes du mot en question afin de proposer ses formes fléchies. Habituellement, cela génère un bruit plus ou moins important dans les contextes récoltés. Par exemple, la recherche de *scor* en corpus

de vulcanologie pourrait donner *scories* et *scorie*, mais aussi *score* qui est moins intéressant. En outre, les expressions polylexicales sont graphiquement respectées (recherche telle quelle) ce qui peut affecter la qualité des contextes obtenus. Par exemple, si l'on cherche *volcan bouclier*, ensuite *volcans boucliers* sur Linguee, les résultats seront différents.

Parfois, les concordanciers bilingues sont critiqués pour la qualité des contextes alignés qu'ils proposent. La plupart du temps, l'alignement est de type 1-1, c'est-à-dire qu'un contexte source est aligné avec un seul contexte cible. Dans certains cas, les contextes ne sont pas des traductions réciproques ou l'alignement n'est pas satisfaisant. Il serait alors intéressant de proposer à l'utilisateur d'autres alignements possibles (de type 1-n). Par ailleurs, les contextes proposés manquent de variétés linguistiques (ex. collocations) et d'informations provenant de domaines de spécialité. En pratique l'utilisation des concordanciers en traduction spécialisée nécessite la sollicitation de ressources complémentaires telles que des dictionnaires, corpus, etc.

## 7.4 Concordanciers bilingues en révision

Les concordanciers bilingues font partie de l'outillage du traducteur en domaine de spécialité. Ils exploitent principalement les corpus parallèles dans lesquels les phrases sources sont préalablement alignées avec des phrases cibles équivalentes. Ces ressources permettent pour un terme saisi par l'utilisateur de consulter ses usages dans les différents contextes retournés par l'outil.

### 7.4.1 Utilisation des concordanciers bilingues en révision

Dans un exercice de traduction, les concordanciers bilingues fournissent au traducteur toutes les occurrences d'un terme à traduire, dans une fenêtre de mots en langue source, ainsi que les segments correspondant en langue cible. Ceux-ci contiennent éventuellement la traduction du terme souhaité. En pratique, l'objectif du traducteur consiste à repérer la bonne traduction du terme saisi dans la concordance proposée en langue cible, afin de sélectionner la meilleure traduction. La figure 7.1 présente TransSearch [1], un exemple de concordancier bilingue proposé par Bourdaillet et al. (2010) en vue d'aider à la traduction à partir de corpus parallèles.

En revanche, lors d'un exercice de révision, de l'anglais vers le français par exemple, le réviseur doit saisir le terme et sa traduction dans le concordancier bilingue mais de façon indépendante en suivant la procédure suivante :

1. http://www.tsrali.com/Main.aspx?cc=true

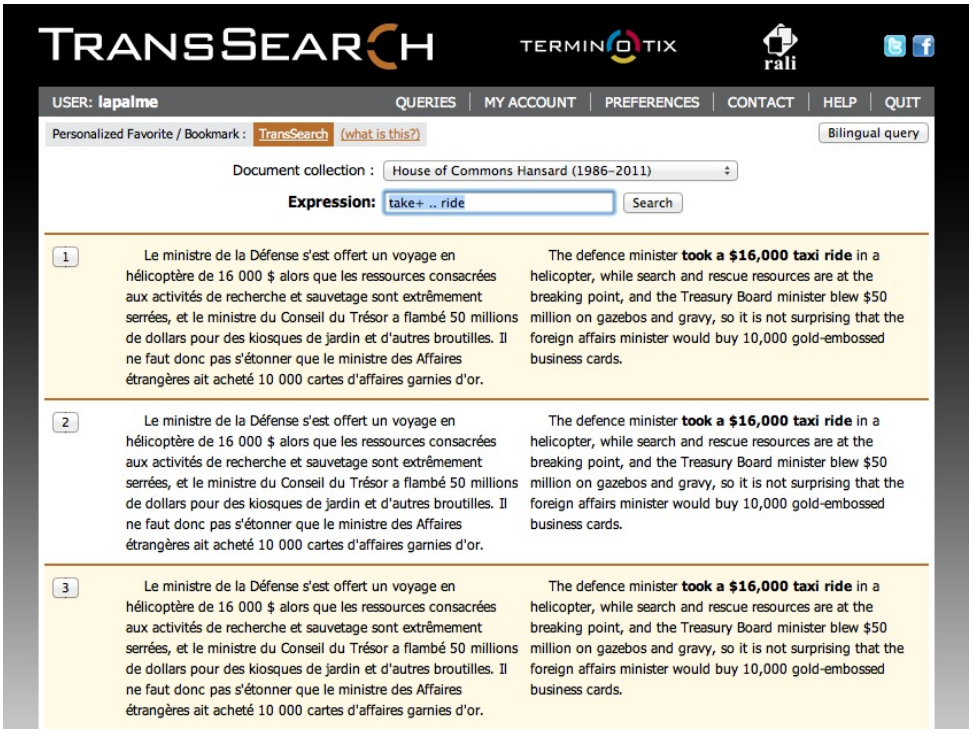


FIGURE 7.1 – *TransSearch, un exemple de concordancier bilingue.*

1. dans un premier temps, il saisit le terme à traduire pour consulter ses occurrences en EN ainsi que leurs équivalentes en FR ;
2. dans un second temps, il saisit la traduction pour consulter ses occurrences en FR ainsi que leurs équivalentes en EN.

Le lien entre le terme saisi et sa traduction s'établit à travers les contextes alignés dans le corpus parallèle. L'un des concordanciers bilingues les plus sollicités tout au long du processus de traduction est Linguee [2]. Les figures 7.2 et 7.3 illustrent l'utilisation de Linguee pour réviser la traduction *blob/goutte*.

Bien que l'intérêt de ces ressources soit indéniable, une limite tient au fait que les corpus parallèles sont rares, notamment en domaines de spécialité. Les contextes proposés par ces outils, sont la plupart du temps généraux et manquent de connaissances spécifiques que l'on peut trouver dans des corpus spécialisés (*cf.* figure 7.2). Ils sont donc appréciés par les lexicographes (Bowker et Barlow 2004, Bourdaillet et al. 2010). Dans la continuité des recherches de Kilgarriff, SketchEngine [3] exploite aussi de grands corpus comparables en supplément des corpus parallèles et des dictionnaires. Cet outil fournit pour un terme source des traductions candidates illustrées par des contextes extraits grâce à des collocations traduites (Baisa et al. 2014), sans toutefois proposer des contextes sources et sans traiter des corpus spécialisés. La démarche suivie par Baisa et al. (2014) consiste à :

1. saisir tout d'abord un mot ;

2. www.linguee.com
3. www.sketchengine.co.uk/bilingual-word-sketch/


FIGURE 7.2 – *Exemple d'application d'un concordancier bilingue sur le terme blob*

2. trouver ses collocatifs en langue source ;
3. traduire le meilleur collocatif (selon la mesure utilisée) avec un dictionnaire bilingue ;
4. considérer cette traduction comme la base d'une collocation pour chercher ses collocatifs en langue cible ;
5. proposer les contextes contenant les collocations en langue cible.

### 7.4.2 Objectifs en révision bilingue

En révision bilingue, dite aussi comparative, le réviseur effectue des va-et-vient entre les ressources sources et cibles, dans un processus répétitif, afin de s'assurer que la traduction est valide dans sa langue cible. Ici, le réviseur, en consultant des segments de textes dans les deux langues, estime quels seront les ponts de transition vers l'autre langue. Les ressources sont consultées de façon bilingue. C'est dans ce cadre que nous envisageons de proposer un prototype de concordancier bilingue permettant de :

1. saisir en même temps un terme et sa traduction attestée afin d'éviter la double utilisation de l'outil et ainsi optimiser le travail du réviseur ;


FIGURE 7.3 – *Exemple d'application d'un concordancier bilingue sur le terme goutte*

2. illustrer le terme et sa traduction par des contextes qui permettront de confirmer ou infirmer la traduction choisie. Ces contextes devront contenir des connaissances linguistiques et conceptuelles spécifiques au domaine en question à partir de corpus comparables. Nous parlons plus précisément des contextes riches en connaissances présentés dans la partie CRC monolingues. Dans le présent chapitre, ces CRC seront exploités de façon bilingue en vue de confirmer (ou infirmer) la traduction.

Un des problèmes mis en évidence dans les expérimentations du chapitre 6, est le faible rappel des CRC conceptuels (fournis pas les patrons de connaissances). Il serait alors trop restreint d'aligner ceux-ci. Nous exploitons les CRC obtenus plutôt par les collocations pour la suite de ce notre travail.

La stratégie sur laquelle se base le concordancier bilingue que nous proposons contient deux étapes :

1. alignement des collocations : nous faisons appel à l'alignement des collocations plutôt qu'à alignement des phrases (il est en effet peu probable de trouver des phrases en correspondance de traduction dans des corpus comparables spécialisés). Ici, les collocations sont considérées comme le seul point d'ancrage assurant à la fois l'iden-

tification bilingue des CRC et une première transition d'un contexte source à un contexte cible.

Par exemple, pour la traduction (*lava, lave*), deux collocatifs peuvent être alignés : *basaltic* et *basaltique*. Les deux collocations *basaltic lava* et *lave basaltique* permettent de récolter des CRC sources et cibles non alignés.

2. alignement des CRC fournis par les collocations retenues dans la précédente étape : nous associons aux CRC sources des CRC cibles équivalents pour une traduction donnée (terme source/traduction attestée). La démarche qui sera suivie consiste à filtrer tout d'abord les contextes de façon monolingue afin d'aligner par la suite les contextes sources avec les cibles.

   Cette étape permettra de proposer pour la traduction (*lava, lave*), les deux contextes « *Shield cones are broad, slightly domed volcanoes built primarily of fluid,* ***basaltic lava*** », et « *Volcan bouclier, volcan de forme ovale, très aplati, dû à l accumulation de coulées de* ***lave basaltique*** *fluide* ».

## 7.5 Alignement de collocations

Nous motivons dans cette section les hypothèses d'alignement des collocations et décrivons leur mise en œuvre. Ces collocations alignées permettront de fournir des CRC dans les langues source et cible.

### 7.5.1 Exploitation interlingue des collocations

Notre objectif consiste à déterminer des « propriétés » invariantes entre les deux langues, permettant au réviseur de construire des « ponts de transition » afin de passer d'un contexte source à un contexte cible équivalent, et de vérifier la traduction en question. Concrètement dans le processus de traduction, la traduction littérale présente un pont entre le texte de départ et le texte final. Cette étape intermédiaire permet de désambiguïser un segment de texte linguistiquement ou cognitivement complexe. Parmi les problèmes auxquels est confronté le traducteur qui ne possède pas une connaissance approfondie de la terminologie, figurent avant tout les termes spécialisés ainsi que leur usage. Ce problème a été abordé par Mammino (1995, p. 7-9)[4] :

> « *Très souvent, l'usage est normalisé. On utilise des termes bien précis pour des aspects et des phénomènes bien particuliers, à un point tel que l'usage devient presque la règle. [...] S'opposer à l'usage, c'est souvent introduire une erreur.* »

4. « *Molto spesso l'uso è standardizzato : si impiegano ben precisi termini per ben specifici aspetti o fenomeni, al punto che l'uso diviene quasi la regola. ... Opporsi all'uso significa opporsi a una scelta che è stata fatta tenendo conto di tutte le circostanze legate alla particolare informazione ; di conseguenza, opporsi all'uso significa spesso introdurre errori.* »

En effet, il est possible que les outils de recherche terminologique à la disposition des traducteurs n'indiquent pas toutes les unités terminologiques en usage dans leurs contextes typiques. La traduction qui ne respecte pas les collocations standards du domaine risque d'être perçue de manière négative par les destinataires experts (Musacchio et Palumbo 2008). Disposant déjà du couple *terme source/terme cible*, il est nécessaire de vérifier la validité de la traduction en fonction des voisinages typiques des termes en question, à savoir leurs collocations. D'autant plus que ces dernières sont considérées comme un « indice » de richesse linguistique marquant les contextes dans lesquels elles apparaissent.

Dans le cadre du concordancier bilingue que nous proposons, nous identifions des équivalents bilingues pour les CRC à partir de corpus comparables, en faisant appel, dans un premier temps, à l'alignement des collocatifs plutôt qu'à l'alignement des phrases. Ici, nous considérons les collocations comme étant un point d'ancrage préservé de façon interlingue pour un couple donné (*terme à traduire/traduction attestée*). Autrement dit, nous postulons que la traduction d'une collocation est une collocation en corpus comparable en domaine de spécialité.

**Hypothèse 1 :** « les collocations sont préservés entre la langue source et la langue cible »

### 7.5.2 Collocations et traduction littérale

Habituellement, le traducteur doit chercher des approximations de la collocation de la langue source vers la langue cible. La traduction littérale est de loin la plus fréquente, d'autant plus qu'elle représente la première disposition du traducteur : si la traduction littérale est juste, il serait mal avisé de tenter à tout prix de l'éviter, car elle peut permettre des équivalences référentielles et pragmatiques (Newmark 1988, p. 68-96). La traduction littérale représente la norme de traduction en traduction spécialisée. En règle générale, plus le texte est technique, plus littérale sera la traduction (Permentiers et al. 1996, p. 64).

Du fait de la sémantique particulière des collocations, notamment en langue générale, leur alignement peut s'avérer problématique. En effet, elles ne peuvent être traduites littéralement, sur la seule base des lexies qui les composent. Par exemple, la traduction mot-à-mot de la collocation *fierce battle* (EN) donne *bataille féroce* (FR) au lieu de *bataille acharnée* (traduction valide). Dans le présent travail, notre objectif n'est pas de traduire les collocations. Il consiste plutôt à proposer des collocations « proches », acceptables pour une révision humaine et aidant à vérifier que la traduction attestée se manifeste dans son contexte d'usage typique. Même si la collocation traduite littéralement ne correspond pas à la meilleure traduction, nous supposons qu'elle permettra au réviseur de s'approcher du sens et de l'usage du terme souhaité à partir de différents contextes, et ainsi d'assurer la cohérence de la traduction produite par le traducteur.

**Hypothèse 2 :** « la traduction littérale d'un collocatif source est un collo-

catif permettant au lecteur de s'approcher du sens et de l'usage du terme cible en question »

### 7.5.3 Alignement des collocatifs

De nombreux travaux tels que ceux de Sharoff (2006) ont été proposés pour identifier automatiquement les équivalents des collocations à partir de corpus comparables. Ces travaux montrent par exemple que les deux mots d'une collocation fournissent un contexte qui facilite la recherche d'associations similaires dans un corpus comparable, *via* la construction de classes grammaticales de termes similaires en corpus. Ici, nous nous appuyons sur l'étiquetage morpho-syntaxique des corpus étudiés afin de repérer les équivalents « proches » d'une collocation source, dans une langue cible. Nous supposons que la catégorie morpho-syntaxique des collocations est identique entre les langues source et cible.

**Hypothèse 3 :** « les catégories grammaticales d'une collocation source et de celle de son équivalent en langue cible sont identiques »

### 7.5.4 Synthèse

L'idée consiste à proposer tout d'abord pour le couple (*terme source/terme cible*), des collocations alignées automatiquement à l'aide d'un dictionnaire bilingue en langue générale.

1. **Extraction de CRC monolingue :** ici, nous avons repris les CRC extraits dans le chapitre 6.
2. **Alignement de CRC :** ces collocations sont ensuite alignées à l'aide d'un dictionnaire bilingue sur la base du collocatif (puisque nous disposons de l'association *terme source/terme cible*).

Les hypothèses sur lesquelles se base notre méthode d'alignement de collocations sont fondées en quelques sortes sur la traduction compositionnelle des collocations en corpus comparables spécialisés. Dans notre cas, le terme et sa traduction sont considérés comme un pivot, une unité alignée en amont. Les travaux de Morin et Daille (2009) ont exploité la compositionnalité pour générer des traductions candidates des termes polylexicaux, notamment les termes complexes.

## 7.6 Alignement de CRC

Après avoir identifié des CRC sources et cibles en s'appuyant sur des couples de collocations traduites, notre objectif consiste maintenant à associer à chaque CRC source des CRC cibles équivalents. Pour cela, nous filtrons tout d'abord les CRC, ensuite nous alignons ceux qui sont retenus, pour un couple de collocations alignées.

### 7.6.1 Filtrage monolingue des contextes

Les phrases non valides récoltées par les collocations, affecteront négativement l'alignement des CRC. Nous mettons en œuvre des critères, que nous qualifions de négatifs, aidant à éliminer les phrases les moins intéressantes. Nous ne retenons que des CRC « normalisés » selon les critères suivants appliqués de façon monolingue :

1. **longueur du contexte :** nous postulons que les phrases courtes ne contiennent pas assez de connaissances autres que la collocation en question. D'autre part, celles qui sont très longues sont difficiles à consulter, et risquent d'illustrer des informations inutiles pour la révision. Seulement les phrases contenant entre 10 et 20 mots pleins, sont retenues. Ce critère a été également adopté par Kilgarriff et al. (2008) ;
2. **présence de pronoms :** dans leur travaux, Kilgarriff et al. (2008) pénalisent les phrases contenant des anaphores pronominales, puisqu'elles présentent un facteur d'ambiguïté. En particulier, les pronoms en début de phrase posent un problème de référence. Nous considérons ceux qui apparaissent au milieu comme moins problématiques d'autant plus qu'ils peuvent référer à des noms dans la même phrase. Dans exemple, « *Desmarest commence ses recherches en volcanologie en 1763, en Auvergne, où* ***il*** *étudie les colonnes de basalte :* ***il*** *est le premier à reconnaître leur origine volcanique* » le pronom *il* ne pose pas de problème en référant au sujet *Demarest* dans la même phrase. Ici, nous choisissons d'éliminer seulement les contextes commençant par un pronom ;
3. **phrases affirmatives :** Kilgarriff et al. (2008) considèrent les phrases interrogatives comme non intéressantes et privilègent plutôt les phrases affirmatives. Nous retenons aussi ce critère pour filtrer les contextes obtenus ;
4. **complexité du contexte :** ce critère qui renseigne sur la lisibilité de la phrase a été également pris en compte par Didakowski et al. (2012). Nous suivons la même stratégie en utilisant un analyseur en dépendances pour éliminer les phrases complexes. Dans notre cas, nous exploitons la somme des scores des arbres syntaxiques possibles d'une phrase donnée pour mesurer sa complexité : plus la phrase est complexe, plus grande est la somme de tous ses arbres possibles.

### 7.6.2 Alignement des contextes

Les CRC obtenus à cette étape ne sont alignés que sur la base des (termes à traduire/traductions) ainsi que leurs collocations alignées. Nous envisageons alors de présenter de façon plus exploitable les CRC, alignés en fonction d'autres points d'ancrage en plus des collocations. Dans les corpus comparables, les phrases parallèles ou lexicalement similaires sont rares. Il serait encore plus restreint d'aligner les CRC sur le lexique qui les compose. Nous proposons par conséquent d'avoir recours à des critères de ressemblance faisant correspondre à un CRC source un CRC cible

équivalent. Ces critères représentent des ponts « statiques » de transition d'une langue à l'autre :

1. **nombre de cognats**[5] **:** les cognats représentent des ponts de transition aisément repérés par le lecteur dans les couples de contextes (source et cible). Les contextes partageant au moins un cognat seront alignés ;
2. **nombre de termes simples traduits :** bien qu'elles soient rares dans le corpus, les phrases contenant des termes traduits sont exceptionnellement utiles pour le réviseur. Les termes simples des corpus étudiés ont été préalablement extraits par un outil d'extraction terminologique. Les contextes contenant au moins un terme et sa traduction, seront alignés.

## 7.7 Conclusion

Après avoir étudié les CRC monolingues dans les chapitres 5 et 6, nous nous sommes intéressés à la problématique d'aider à la révision d'une traduction (terme à traduire/traduction attestée) afin d'assurer la cohérence et la fiabilité du document produit par le traducteur. Nous avons proposé des couples de contextes riches en connaissances alignés en corpus comparables spécialisés. Dans un premier temps, nous avons fait appel à l'alignement des collocations pour extraire des CRC potentiellement proches. Dans un second temps, nous avons aligné ces CRC en utilisant des filtres monolingues ainsi que des critères bilingues d'alignement. Dans ce travail d'alignement, nous avons exploité les collocations, les cognats ainsi que les termes simples comme étant des points d'ancrage suffisamment fiables pour extraire des CRC bilingues en corpus comparables spécialisés.

5. Pour Léon (2008), deux cognats sont deux mots ayant les mêmes 4 premières lettres.

# Évaluation

8

Sommaire

## 8.1 Introduction

L'évaluation des CRC est une problématique qui a été au cœur de nos travaux. En effet, tous les travaux de l'état de l'art (Kilgarriff et al. 2008, Schumann 2011; 2012b;a, Didakowski et al. 2012, Marshman 2014) quels qu'ils soient dans un cadre terminologique, lexicographique ou encore informatique, ont recours à une évaluation manuelle qui reste sans doute l'un des exercices les plus délicats et coûteux en termes logistique. À défaut de corpus étalon de CRC ou d'une méthode de référence, nous avons été contraint de suivre la même stratégie : évaluer manuellement notre méthode. L'inconvénient de cette évaluation est qu'elle peut être perçue comme plus ou moins subjective puisqu'elle dépend le plus souvent de l'interprétation humaine. Cette dernière étant sujette à différents facteurs tels que les prés-requis de la personne dans le domaine en question, sa maîtrise de la langue (natif ou non natif), le temps consacré à l'exercice, ou même les conditions ergonomiques de l'évaluation. Néanmoins, elle permet d'étudier au plus prêt les phénomènes linguistiques que peuvent révéler ces CRC, et ainsi, de confirmer ou infirmer nos hypothèses et nos observations. Du fait que nous abordons l'extraction des CRC en vue d'aider à la traduction spécialisée, l'évaluation devrait porter sur cette notion de « riche en connaissances » et qui est fortement liée à l'utilité des contextes identifiés, dans un exercice de traduction. Nous avons alors réalisé des expériences avec des réviseurs, mettant en œuvre les CRC bilingues proposés dans un exercice réel du métier du traducteur au quotidien.

## 8.2 Évaluation manuelle

Nous présentons dans cette section les ressources supplémentaires par rapport à celles présentées dans le chapitre 6. Nous détaillons l'évaluation manuelle réalisée ainsi que les résultats obtenus.

### 8.2.1 Dictionnaire bilingue

Nous avons utilisé pour l'alignement automatique des collocations, un dictionnaire bilingue de langue générale anglais-français contenant 145 542 entrées : ELRA [1]. Ce dictionnaire indique également l'étiquette morpho-syntaxique de chaque entrée.

### 8.2.2 Liste terminologique d'évaluation

Les données d'évaluation portent sur les termes simples provenant des expérimentations effectuées dans la partie concernant les CRC monolingues, à savoir 29 termes pour chaque corpus (*cf.* 6.4). Les termes complexes ont été écartés pour deux raisons : d'une part, en exploitant les col-

1. http://catalog.elra.info/product_info.php?products_id=666

locations, nous traitons indirectement d'autres expressions polylexicales, notamment les termes complexes qui partagent des critères (*ex*. la catégorie syntaxique et la fréquence) avec les collocations recherchées. D'autre part, l'accès à des termes simples et à leurs contextes, donnent également accès à certains termes complexes (ainsi qu'à leurs contextes) constitués à partir de ces termes simples.

Les termes ont été séparés en sources et cibles correspondant ainsi à des couples de traduction attestée auxquels seront associés des CRC dans les deux langues. Nous disposons de deux listes de 29 termes illustrés par la table 8.1.

| **Termes sources (EN)** | **Traduction attestée (FR)** |
|---|---|
| **Vulcanologie** | |
| basalt | basalte |
| cinder | scorie |
| crater | cratère |
| cone | cône |
| debris | débris |
| dome | dôme |
| fountain | fontaine |
| gas | gaz |
| lava | lave |
| magma | magma |
| scoria | scorie |
| tephra | téphra |
| volcan | volcan |
| vesicle | vésicule |
| eruption | éruption |
| **Cancer du sein** | |
| curage | dissection |
| dépistage | screening |
| ganglion | node |
| carcinome | carcinoma |
| séquelle | relapse |
| rechute | relapse |
| envahissement | involvement |
| douleur | distress |
| zonectomie | lumpectomy |
| surdosage | boost |
| récidive | recurrence |
| exérèse | excision |
| morbidité | morbidity |
| guérison | recovery |
| anomalie | abnormality |

TABLE 8.1 – *Couples de termes de référence*

### 8.2.3 Protocole d'évaluation et consignes d'annotation

#### Protocole d'évaluation

La démarche lors de nos expérimentations a été de :

1. évaluer, dans un premier temps du point de vue monolingue, les contextes fournis par les collocations alignées automatiquement. En pratique, il s'agit de la même évaluation monolingue que celle réalisée dans le chapitre 6, avec un filtre supplémentaire sur les collocations (retenir seulement celles qui ont des équivalents).
2. évaluer parallèlement ces couples de CRC sources et cibles, correspondant ainsi à une évaluation bilingue.

La table 8.2 détaille les quatre couples de contextes possibles, nous nous intéressons en particulier au *couple collocations 4*. Dans les autres cas, l'alignement des contextes serait moins utile pour la révision.

| **Contextes sources** | **Couples de collocations** | **Contextes cibles** |
|---|---|---|
| non CRC | couple collocations 1 | non CRC |
| non CRC | couple collocations 2 | CRC |
| CRC | couple collocations 3 | non CRC |
| **CRC** | **Couple Collocations 4** | **CRC** |

TABLE 8.2 – *Types de contextes selon l'alignement des collocations*

| **Terme** | **Collocatif** | **Phrases** | **Évaluation** |
|---|---|---|---|
| Lava (EN) | gush | (a) ***Lava*** *began* ***gushing*** *out of mayon's crater before dawn yesterday, accompanied by loud rumblings.* | valide |
| Lava (EN) | gush | (b) *The weight of* ***lava*** *above caused* ***lava*** *to* ***gush*** *freely from the vents, and for a few hellish hours rock flowing almost like water engulfed villagers, their livestock, and wild elephants on the slopes.* | valide |
| Lave (FR) | jaillit | (c) *Profitant de la cassure, la* ***lave jaillit****.* | valide |
| Lave (FR) | jaillit | (d) *De grandioses fontaines de* ***laves*** *ont* ***jailli*** *au travers des fissures qui se sont formées sur la face nord du dolomieu, le cône central de ce volcan culminant à près de 2 500 mètres.* | non valide |

TABLE 8.3 – *Exemple de validation monolingue des contextes pour (lava, gush) et (lave, jaillir)*

**Consignes d'annotation**

Nous avons évalué manuellement les contextes résultant sans faire intervenir des annotateurs externes. Certains cas problématiques (*cf.* section 8.2.5) ont été discutés afin de prendre des décisions cohérentes. Les critères que nous avons adoptés pour annoter les contextes bilingues de chaque couple (terme source/traduction attestée) sont les suivantes :

1. **valider tout d'abord les contextes du point de vue monolingue :** déterminer pour chaque terme source et sa traduction si les hypothèses (les collocations sont préservées en corpus comparables, et leurs traductions compositionnelles peuvent être linguistiquement intéressantes) qui ont donné lieu à notre méthode d'alignement sont bien présentes dans les contextes ; ensuite évaluer les contextes sources et cibles indépendamment.

   Toutes les traductions des collocations sont acceptables. Pour la validation manuelle, un concordancier [2] monolingue et un dictionnaire bilingue [3] ont été utilisés pour vérifier que les collocatifs sont alignés en respectant les catégories syntaxiques (hypothèse 3) et ainsi identifier les erreurs d'étiquetage. Pour la traduction $(lava(EN), lava(FR))$ produite par le traducteur, deux collocatifs verbaux *gush* (dans *lava gush*) et *jaillir* (dans *lave jaillit*) ont été alignés, donnant les contextes (a), (b), (c) et (d) (*cf.* table 8.3).

   Les mêmes consignes de la section 6.4.1 ont été respectées : nous avons récupéré les résultats des évaluations de CRC monolingues. Par exemple, pour le terme *lava* $(EN)$, la collocation (*lava*, *gush*) dont le collocatif *gush* est aligné avec *jaillir* qui forme une collocation avec *lave* $(FR)$ (*cf.* table 8.3). Le dernier contexte (d) de la table 8.3 n'a pas été validé car le collocatif *jaillir* modifie le terme *fontaines de laves* plutôt que *lave*. Les autres contextes (a), (b) et (c) ont été annotés comme CRC.

2. **valider ensuite les contextes alignés :** vérifier si les contextes alignés fournis pour chaque couple de collocations retenues dans 1) sont valides du point de vue bilingue. Un contexte bilingue n'est validé que si les deux contextes sources et cibles sont des CRC monolingues, c'est-à-dire qu'ils ont été initialement considérés comme CRCC ou CRCL.

   Pour valider manuellement un couple de CRC, il faut que les critères d'alignement soient également valides dans une fenêtre de 7 mots (approximativement) contenant le terme en question ou sa traduction attestée. Une première étude des contextes obtenus a montré que, dans certains cas, un nombre très élevé de candidats est proposé du fait que notre méthode permet d'aligner le même contexte source avec plusieurs contextes cibles. D'une part, consulter chaque contexte entièrement rend très laborieux l'exercice de l'évaluation. D'autre part, même s'ils respectent les critères d'alignement, plusieurs contextes bilingues peuvent s'avérer « non intéressants » du

2. http://www4.caes.hku.hk/vocabulary/concordancer.htm
3. http://www.wordreference.com/

point du vue linguistique, dans le sens où ils ne renferment pas de véritables points d'ancrage : dans ce cas, les critères d'alignement ne serviront pas de pont de transition pour passer d'un contexte à l'autre dans un exercice de révision. Nous présentons des exemples dans la section 8.2.5. Une solution pratique pour palier ces problèmes d'évaluation, est de valider les contextes, dans une fenêtre de mots aisément consultable, si les critères d'alignement peuvent être perçus comme des propriétés linguistiques partagées par les deux contextes en question. Par exemple :

— traduction à réviser : *lava, lave*

— CRC source : *Shield cones are broad, slightly domed volcanoes built primarily of fluid,* ***basaltic lava****.*

— CRC cible : *Volcan bouclier, volcan de forme ovale, très aplati, dû à l'accumulation de coulées de* ***lave basaltique*** *fluide.*

Il s'agit d'un exemple de CRC fourni par le concordancier bilingue proposé en corpus comparable spécialisé. La traduction à réviser étant *lava, lave* les collocatifs *basaltic, basaltique* ont tout d'abord été extraits automatiquement, et ensuite traduits en utilisant le dictionnaire bilingue ELRA. Les cognats repérés dans ces CRC sont *volcanoes* et *volcan*. Les termes candidats préalablement identifiés par TermSuite et traduits par la suite par le même dictionnaire, sont *shield, fluid, bouclier* et *fluide*.

Dans certains CRC proposés, les critères d'alignement peuvent s'avérer non intéressants du point de vue linguistique. Cependant, ces CRC peuvent également renfermer d'autres unités lexicales qui n'ont pas pas été traduites et qui sont en relation sémantique dans les deux langues : des synonymes, des hyperonymes, des méronymes, autres collocations ou autres variantes du terme, etc. Ces éléments linguistiquement intéressants attireraient potentiellement l'attention des réviseurs. Il serait alors maladroit d'ignorer ces contextes en les considérant comme invalides ou non intéressants. Nous évoquerons ces situations parmi les difficultés rencontrées dans la section 8.2.5. Nous nous contentons de vérifier dans une fenêtre de 7 mots (approximativement) s'il existe des éléments « proches » sur le plan lexical ou sémantique. Ceux-ci peuvent également être considérés comme un pont de transition entre les contextes. Dans l'exemple suivant, les CRC ont été validés grâce à la paire de mots (*ejection, projettent*) qui modifient *incandescent cinder* et *scories incandescents* et qui n'ont pas été traduits automatiquement :

— traduction à réviser : *cinder, scorie*

— collocations alignées : *(cinder, incandescent)* et *(scorie, incandescent)*

— CRC source : *Strombolian eruptions are named for Stromboli volcano off the west coast of Italy, where a typical eruption consist of the rhythmic ejection of* ***incandescent cinder****, lapilli, and bombs to heights of a few tens or hundreds of feet meters.*

— CRC cible : *Le dynamisme strombolien s'exprime par des explosions rythmiques qui projettent des blocs et des* ***scories incandescents****.*

### 8.2.4 Résultats

Nous avons appliqué la méthode proposée dans le chapitre 6 sur les deux listes de la table 8.1. Nous tenons à préciser que l'évaluation effectuée porte sur les aspects qualitatif et quantitatif des contextes alignés fournis après avoir appliqué l'alignement des collocations (section 7.5) ainsi que l'alignement des contextes (section 7.6). Il s'agit d'une évaluation globale de la méthode proposée.

En ce qui concerne l'évaluation bilingue des contextes alignés, deux expérimentations ont été effectuées : alignement des contextes avec et sans filtres. Ici, notre but n'est pas d'évaluer ou de valider les critères de filtrage proposés, mais plutôt d'étudier leur impact sur la qualité des contextes alignés. Ainsi, nous pourrons définir par la suite une stratégie permettant de proposer des CRC bilingues à partir de corpus comparables.

| corpus | # termes | # termes alignés | # couples de coll. alignées | # phrases | # couples de CRC alignés | P. couples de CRC alignés valides |
|---|---|---|---|---|---|---|
| **Sans filtres** | | | | | | |
| Vulcano EN | 15 | 10 | 23 | 677 | 309 | 43,04 % |
| Vulcano FR | 14 | | | 665 | | |
| **Avec filtres** | | | | | | |
| Vulcano EN | 15 | 10 | 16 | 241 | 157 | 61 % |
| Vulcano FR | 14 | | | 296 | | |

TABLE 8.4 – *Évaluation bilingue des CRC bilingues : alignement de CRC avec et sans filtres appliqués*

| corpus | # termes | # termes alignés | # couples de coll. alignées | # phrases | # couples de CRC alignés | P. couples de CRC alignés valides |
|---|---|---|---|---|---|---|
| **Sans filtres** | | | | | | |
| Cancer EN | 15 | 8 | 22 | 229 | 7 279 | 48,38 % |
| Cancer FR | 14 | | | 604 | | |
| **Avec filtres** | | | | | | |
| Cancer EN | 15 | 7 | 7 | 95 | 1 007 | 61,90 % |
| Cancer FR | 14 | | | 190 | | |

TABLE 8.5 – *Évaluation bilingue des CRC bilingues : alignement de CRC avec et sans filtres appliqués*

Les table 8.4 et 8.5 illustrent respectivement l'analyse des CRC bilingue (avec et sans filtres) en vulcanologie et en cancer du sein. # *Termes alignés* est le nombre de couples (terme/traduction attestée) ayant des CRC sources et cibles alignés. Dans la table 8.4, 10 traductions parmi les 15 ont obtenu au moins un couple de CRC alignés. # *couples de coll. alignées* représente le nombre de couples de collocations dont le collocatif est traduit par un dictionnaire bilingue. La colonne # *phrases* est le nombre de phrases fournies par les collocations alignées. Dans la même table 8.4, sans filtres, les 23 couples de collocations fournissent 677 contextes candidats en EN et 665 en FR. # *couples de CRC alignés* contient le nombre de couples (CRC source, CRC cible) pour tous les termes alignés (# *Termes alignés*).

Ces tables indiquent que le nombre de couples de traductions attestées (# *termes alignés*) est assez bien conservé même si cinq couples de traductions en vulcanologie et huit couples de traductions en cancer du sein n'ont pas été couverts par notre méthode : *basalte/basalte*, *volcan/volcan*, *fountain/fontain*, *tephra/téphra* et *vesicle/vésicule*. Nous expliquons ceci par le

nombre réduit en plus de la nature non variée des collocations extraites, comme pour *vésicule* qui n'apparaît que deux fois dans le corpus. L'analyse des colonnes # *phrases* par rapport à # *couples de collocations alignées*, pour les deux corpus, montre que les collocations alignées sont productives : chaque couple de collocations produit en moyenne 28 contextes sans filtre et 15 avec filtres, pour chaque langue en vulcanologie. Nous constatons que même si l'application des filtres réduit le nombre de CRC alignés, elle améliore significativement la précision des critères d'alignement puisqu'elle passe de 43 à 61 % en vulcanologie et de 48 à 62 % en cancer du sein.

### 8.2.5 Difficultés rencontrées

Les principales difficultés rencontrées portent sur l'évaluation. Nous abordons les cas les plus significatifs, ainsi que les solutions que nous avons adoptées :

#### Nombre de CRC candidats très élevé

7 279 couples de phrases (sans filtre) ont été proposés pour le corpus cancer du sein. Cela implique une double vérification (plus de 14 000 phrases à consulter), ainsi que plusieurs allers-retours entre les contextes avant de décider l'annotation. Actuellement, l'évaluation intégrale des sorties de notre méthode est en hors de notre portée. Ainsi, nous nous sommes contentés d'évaluer les 30 premiers contextes bilingues pour chaque couple de collocations. Si le même contexte source est aligné avec plusieurs contextes cibles, nous avons évalué seulement les cinq premiers. Concernant le corpus de vulcanologie, le nombre de contextes étant plus raisonnable (307 pour vulcanologie EN), tous les candidats ont été évalués.

#### Contextes bilingues pauvres

Les contextes renferment des termes simples traduits automatiquement par le dictionnaire, ou des cognats. Bien qu'ils soient intéressants du point du vue terminologique et lexical, les contextes contenant ces éléments peuvent s'avérer moins utiles comme dans les exemples a) et b) et d) de la table 8.6. Une des solutions que nous avons adoptée consiste à évaluer les contextes sur la base des éléments présents dans le voisinage des termes en question, dans une fenêtre de sept mots approximativement. Par exemple, dans les contextes c) et e) les expressions (*consist primarily of*, *se caractérise par*) et (*made of*, *créent de*) peuvent être interprétées comme des relations ou des « actions » réciproques dans les deux langues, renfermant les termes en question.

**Jugement des contextes**

L'objectif d'aligner les CRC amène à se poser des questions sur les points d'ancrage qui peuvent exister dans les corpus comparables. Nous nous sommes basés sur des « propriétés comparables » de nature terminologique (termes simples traduits) et lexicographique (les cognats). Ici, nous avons postulé que l'intérêt du lecteur (plus particulièrement le réviseur) se focaliserait sur les voisinages du terme en question, ce qui reste dans l'idée traditionnelle des concordanciers. Cependant, la difficulté majeure dans l'évaluation a été de juger la qualité de l'alignement des CRC. Certains CRC sources et cibles peuvent être moins intéressants du point de vue monolingue, mais donnent des alignements de bonne qualité. D'autres sont plutôt plus intéressants en langue source, mais moins utiles du point de vue bilingue.

## 8.3 Évaluation expérimentale en révision

Après avoir évalué manuellement la qualité des CRC bilingues, nous étudions leur apport dans un exercice de révision. Nous mettons alors le concordancier bilingue que nous proposons, désormais appelé KRCTool, en expérimentation.

### 8.3.1 Données expérimentales

Le texte qui a été utilisé pour les expériences avec les traducteurs (section 6.6) a également été repris pour les expérimentations avec les réviseurs. Nous avons retenu l'une des traductions les plus perfectibles produites par les étudiants lors des expériences menées dans le cadre du projet CRISTAL. Des extraits de ces textes sont illustrés en table 8.7. Dans les textes traduits, nous avons remplacé trois termes par des mots issus de la langue générale. Les réviseurs devaient tout d'abord détecter l'anomalie, et en utilisant KRCTool, ils pouvaient ensuite les corriger par des traductions plus spécifiques à la vulcanologie. La table 8.8 contient les termes sources, leurs traductions proposées ainsi que leurs traductions exactes.

Une vue détaillée du raisonnement auquel nous nous attendons est la suivante : lorsque le couple *blob* et *goutte* est cherché dans KRCTool, un seul CRC cible contenant *goutte* sera affiché. Celui-ci, illustre un usage de *goutte* qui ne correspond pas à *blob* en CRC source. Un bon réviseur doit alors désapprouver *goutte* et chercher une solution alternative. L'outil propose d'autres traductions possibles telles que *projection*. En revanche, lorsque *blob* et *projection* sont recherchés dans KRCTool, des CRC de la collocation *projection de lave* sont affichés avec *blob of magma*. Les CRC alignés permettent de confirmer que *projection* est la traduction la plus acceptable de *blob*.

| CRC source | CRC cible | Évaluation |
|---|---|---|
| **a)** *The distance between the lake surface and the lowest point on the* ***crater rim****, known as the freeboard, has shrunk from 8.3 metres 27ft to 4.5 metres in the last two months, according to the oxfam team's senior geologist, professor Kelvin Rodolfo.* | *Une façon de prendre le pouls du colosse au* ***bord du cratère*** *du vulcano îles éoliennes, un chercheur mesure à distance la température d une fumerolle à l'aide d'un pistolet à infrarouges.* | invalide |
| **b)** *Although large eruptions and lava flows are uncommon, smaller eruptions occur very frequently and often hurl blobs of lava above the* ***crater rim****.* | *Au* ***bord du cratère*** *actuel sans doute celui de l'éruption de 1631, vous entendrez sûrement l'un des guides essayer de communiquer un frisson de crainte à un groupe de touristes, en affirmant que le Vésuve va se réveiller très, très bientôt peut-être même dans la demi-heure qui suit.* | invalide |
| **c)** *The eruption will consist primarily of* ***dome growth****, but as with all* ***dome growth****, minor explosive activity is also possible.* | *Cette éruption se caractérise par une activité éruptive comparable à celle du début de 1902, avec* ***croissance d'un dôme*** *dans le cratère sommital et destruction de ce dernier par des explosions dirigées latéralement.* | valide |
| **d)** *A volcano is essentially an opening or a vent through which this magma and the dissolved* ***gases*** *it contains are* ***discharged****.* | *Certains volcans de cet alignements présentent des risques importants car leur lac de cratère a* ***libéré des gaz*** *toxiques : 37 morts au lac monoun en 1984 et 1 746 morts au lac Nyos en 1986.* | invalide |
| **e)** *Lava shields, a shield volcano made of* ***basaltic lava*** *.* | *Au fond du rift, des remontées de* ***lave basaltique*** *créent de nouveaux volcans .* | valide |

TABLE 8.6 – *Exemples d'évaluation et difficultés rencontrées (en gras les collocations en question, souligné : critère d'alignement ; et*

| **Texte à traduire** | |
|---|---|
| Partie 1 | *CINDER CONES*<br>*Basalt **cinder** cones are the most common kind of volcano worldwide. They are also some of the smallest volcanoes. A typical eruption goes through two stages. The first is called the fountaining stage. When the basalt magma first breaks out at the surface, the dissolved gases **bubble off** vigorously enough to carry **blobs** of magma into the air with them. The **blobs** may rise up 2,000 feet or more.* |
| Partie 2 | *During their flight through the air, the **blobs** cool to solid pieces called **cinders** (also called scoria), which are typically no bigger than a softball. These spongelike rocks contain holes, called **vesicles**, where gas bubbles got trapped inside. As the **cinders** fall back to Earth, they form layers that pile up into a cone-shaped hill. The very top, where the magma **spewed out**, usually ends up with a small crater. Once most of the gassy magma has **erupted**, fountaining stops.* |
| **Proposition du traducteur** | |
| Partie 1 | *CONES DE SCORIES*<br>*Les cônes de **débris** basaltiques sont un des types de volcans les plus communs au monde. Ils font également partie des plus petits volcans. Une éruption typique se déroule en deux étapes. La première est appelée étape d'expulsion. Lorsque le magma basaltique explose à la surface, les gaz dissous **partent** de manière si puissante qu'ils emportent avec eux des **boules** de magma. Les **boules** peuvent être expulsés à au moins 600 mètres de hauteur.* |
| Partie 2 | *Pendant leur envol, les **boules** refroidissent et se transforment en morceaux solides appelés **débris**, généralement de la taille d'un ballon de football. Ces roches spongieuses contiennent des cavités appelées **poches** dans lesquelles les bulles de gaz sont retenues. Lorsque les **débris** retombent sur terre, elles forment des couches qui se superposent pour produire un monticule cônique. Le dessus, duquel le magma est **expulsé**, finit généralement par être constitué d'un petit cratère. Une fois que le magma gazeux est **sorti**, l'étape d'expulsion prend fin.* |

TABLE 8.7 – *Le texte et sa traduction proposée par le traducteur, à réviser pendant les expériences. Les termes en gras sont des erreurs ajoutées manuellement.*

| **Terme source** | **Traduction modifiée** | **Traduction exacte** |
|---|---|---|
| cinder | débris | scorie, cendre |
| vesicle | poche | vacuole, vésicule |
| blob | boule | paquet, projection |

TABLE 8.8 – *Termes sources et traductions modifiées*

| **Groupe A** | **Texte 1** | **Texte 2** | **Durée (min)** |
|---|---|---|---|
| Phase 1 : ressources communes | Aa | Ab | 20 |
| Phase 2 : KRCTool | Ab | Aa | 20 |
| **Groupe B** | **Texte 1** | **Texte 2** | **Durée (min)** |
| Phase 1 : KRCTool | Ba | Bb | 20 |
| Phase 2 : ressources communes | Bb | Bb | 20 |

TABLE 8.9 – *Répartitions des groupes*

### 8.3.2 Protocole d'expérimentation

### 8.3.3 Résultats

La table 8.10 résume les interventions des réviseurs avec et sans KRCTool. Nous focalisons nos analyses seulement sur les trois termes en question, à savoir *cinder, vesicle* et *blob*. *Nb* représente le nombre de corrections apportées par tous les réviseurs pour chaque terme. *Total* est le nombre de corrections totales réalisées par réviseur. $x$ désigne une correction apportée. $T2 \geq T1 \geq 1$ permet de déterminer pour un réviseur s'il a apporté le même nombre de corrections avec KRCTool qu'avec les autres outils, ou plus. $1 < T1 < T2$ indique si les ressources habituelles ont strictement été plus efficaces.

En se basant sur la table 8.10, nous pouvons remarquer que chacun de ces termes a été révisé par 1 à 4 étudiants parmi 14. Tous les termes ont été remplacés par des traductions exactes. Nous constatons que le groupe B effectue plus de corrections en se basant sur KRCTool que le groupe A. Ceci est expliqué par le fait que le groupe B a commencé dans la phase 1 avec KRCTool. Cependant, le groupe A utilise tout d'abord d'autres outils en phase 1, puis KRCTool seulement en phase 2 : l'exercice a été majoritairement effectué par le groupe A avec leurs ressources communes. Cela montre que les réviseurs qui ont eu accès tout d'abord à leurs outils habituels s'en sont contentés, et qu'ils étaient moins intéressés par KRCTool. Seulement deux étudiants (Ba1 et Bb7) ont effectué plus de corrections avec KRCTool. La table 8.10 montre également qu'en utilisant KRCTool, quatre étudiants parmi les 13 qui ont apporté des corrections ont accompli les mêmes corrections avec KRCTool, voire mieux. En revanche, cinq étudiants ont effectué de meilleures révision en utilisant des outils communs.

Il ressort de ce travail que le concordancier bilingue que nous proposons n'est pas aussi utile que les outils communs souvent utilisés en traduction spécialisée. Toutefois, lors d'un questionnaire après les expérimentations, certains étudiants ont apprécié le fait que KRCTool fournit des contextes spécialisés auxquels ils n'avaient pas accès en utilisant leurs ressources habituelles. Cet outil pourrait être alors exploité en complément d'autres ressources d'aide à la traductions. Nous devons admettre que ces étudiants ne sont qu'en Master 1 et ne disposent pas de connaissances acquises sur ce domaine spécialisé, pour corriger tous les termes comme le ferait un réviseur professionnel.

## 8.4 Conclusion

Dans ce travail, nous nous sommes intéressés à la problématique d'aider à la révision d'une traduction (terme à traduire/traduction attestée) afin d'assurer la cohérence et la fiabilité du document produit par le traducteur. Nous avons proposé un prototype de concordanciers bilingues : KRCTool. Cet outil accepte en entrée un terme et sa traduction, qui repré-

| Terme | Nb | Aa1 | Aa2 | Ab6 | Ab7 | Ab8 | Ba1 | Ba2 | Ba3 | Ba4 | Bb6 | Bb7 | Bb8 | Bb9 | Bb10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **avec des ressources communes** | | | | | | | | | | | | | | | |
| cinder | 6 | - | - | - | - | x | - | x | x | - | - | x | x | x | - |
| blobs | 3 | - | - | - | - | - | - | - | x | x | - | - | - | x | - |
| vesicles | 4 | x | x | - | - | - | - | x | x | - | - | - | - | - | - |
| **Total (T1)** | | 1 | 1 | - | - | 1 | - | 2 | 3 | 1 | - | 1 | 1 | 2 | - |
| **avec KRCTool** | | | | | | | | | | | | | | | |
| cinder | 2 | - | - | - | - | x | - | x | - | - | - | - | - | - | - |
| blobs | 4 | - | - | - | - | - | - | - | - | x | - | x | x | x | - |
| vesicles | 1 | - | - | - | - | - | - | - | - | - | - | x | - | - | - |
| **Total (T2)** | | - | - | - | - | 1 | - | 1 | - | 1 | - | 2 | 1 | 1 | - |
| **T2 $\geq$ T1 $\geq$1** | | **-** | **-** | **-** | **-** | **x** | **-** | **-** | **-** | **x** | **-** | **x** | **x** | **-** | **-** |
| **1 $<$ T1 $<$ T2** | | **x** | **x** | **-** | **-** | **-** | **-** | **x** | **x** | **-** | **-** | **-** | **-** | **x** | **-** |

Table 8.10 – *Résultats des révisions avec et sans KRCTool (x désigne une correction apportée)*

sente une nouvelle fonctionnalité dans les concordanciers ; et retourne des CRC alignés en corpus comparables spécialisés. KRCTool repose sur une méthodologie exploitant les collocations, les cognats ainsi que les termes simples, traduits avec un dictionnaire de langue générale, comme étant des points d'ancrage favorisant l'identification et l'alignement des CRC. Les premières évaluations et expériences menées auprès de réviseurs ont montré que même si leur apport n'est pas manifeste, les CRC bilingues proposés sont une piste prometteuse dans un exercice d'aide à la révision, malgré la difficulté de celui-ci. Le nombre limité de termes que nous avons utilisé que ce soit dans l'évaluation ou dans les expériences tient au fait que notre étude s'est focalisée sur l'aspect qualitatif des CRC obtenus, ainsi que la complexité de l'évaluation.

# Conclusion générale

# 9

Cette thèse avait pour but d'exploiter les corpus comparables dans le cadre de la traduction assistée par ordinateur. Nos études se basaient plus particulièrement sur la notion de contextes riches en connaissances ainsi que son intérêt en aide à la traduction spécialisée. Ce travail, qui émarge sur différents domaines scientifiques allant de la linguistique de corpus au TAL, s'est focalisé sur l'extraction de connaissances et sur le traitement du multilinguisme, notamment en corpus comparables. En effet, nous avons poursuivi les objectifs suivant : (i) extraction des contextes riches en connaissances (CRC) (Meyer 2001) en corpus monolingues de spécialité en vue d'aider à la compréhension des termes à traduire ou de leurs traductions possibles ; (ii) exploitation des contextes riches en connaissances en corpus comparables spécialisés, qui constituent notre matière première, pour aider à la révision en traduction terminologique ; et (iii) intérêt de ces contextes dans un environnement de traduction assistée par ordinateur.

## 9.1 Contribution

### 9.1.1 Extraction de connaissances en corpus monolingues

Le terme de CRC que l'on doit à Meyer (2001) a été situé dans une évolution des travaux visant la constitution de terminologies. Cette dernière est passée de méthodes s'appuyant uniquement sur les connaissances des experts, à des méthodes qui utilisent les textes comme réservoirs de connaissances pour extraire des définitions. Celles-ci sont validées dans une second temps par les experts. Dans la présente thèse, l'extraction des CRC s'inscrit également dans l'évolution du processus de traduction. En effet, l'analyse du fonctionnement d'un terme à traduire (ou sa traduction potentielle) dans les textes se base sur l'étude des contextes dans lesquels il apparaît. Or, tous les contextes ne sont pas utiles pour le traducteur. Les CRC peuvent être perçus comme des contextes jouant un rôle majeur dans la compréhension (c'est-à-dire la définition) des termes et leur fonctionnement linguistique. Nous avons distingué deux types de CRC : les contextes riches en connaissances conceptuelles et les contextes riches en connaissances linguistiques.

**Contextes riches en connaissances conceptuelles**

Ces contextes sont repérables par des patrons lexico-syntaxiques explicitant des relations conceptuelles entre les termes d'un domaine donné. Nous parlons de patrons de connaissances (PC). Les PC étudiés sont inspirés de nombreuses études dans lesquelles les relations les plus fréquemment étudiées sont l'hyperonymie et la méronymie. L'idée première de l'étude des contextes riches en connaissances conceptuelles a été d'associer systématiquement une interprétation de type sémantique à des patrons de connaissances stables. Dans le cadre du projet CRISTAL, Lefeuvre et Condamines (2015) se sont intéressées au fonctionnement des patrons de connaissances dans différents corpus spécialisés en tenant compte de leur variation : selon le domaine et le genre textuel. Le résultat qui a découlé de cette étude est une liste générique de patrons de connaissances stables présentant deux intérêts majeurs. D'une part, ils limitent le nombre d'occurrences proposées par les outils qui peuvent atteindre plusieurs centaines s'ils ne sont pas filtrés par les patrons de connaissances. D'autre part, ils permettent de comprendre et interpréter aisément les formulations linguistiques, et d'anticiper le fonctionnement des termes dans des situations conceptuelles similaires.

**Contextes riches en connaissances linguistiques**

L'état de l'art (Kilgarriff et al. 2008, Rundell et Kilgarriff 2011, Didakowski et al. 2012) s'inscrit dans l'identification d'exemples lexicographiques introduits sous l'appélation *good dictionary examples*. Même s'ils sont différents des CRC au sens terminologique, les exemples de Kilgarriff et ses collègues contiennent des informations sur les contextes réguliers (telles que les collocations) qui peuvent être utiles en traduction. L'acquisition automatique de ces contextes réguliers s'est principalement focalisée sur des corpus généraux. Toutefois, la nécessité de repérer des régularités concernant le genre textuel ou des domaines de spécialité a été mise en évidence dans plusieurs travaux fondateurs (Halliday 1978, Kittredge 1982, Biber 1991). Les contextes riches en connaissances linguistiques portent sur le fonctionnement distributionnel des termes dans les discours. Il s'agit d'identifier, à partir d'usages réels, les contextes d'apparition privilégiés des termes. Étant des voisinages typiques, notamment en domaine de spécialité, les collocations, qu'ils soient nominaux, adjectivaux ou verbaux, sont particulièrement appréciées par les traducteurs. Nous nous sommes appuyés sur les collocations pour identifier les connaissances linguistiques.

**Apport des CRC monolingues en traduction**

Contrairement aux travaux existants qui se restreignent soit à des patrons de connaissances pour extraire des définitions (Marshman 2014) soit à des collocations afin d'identifier des exemples lexicographiques (Kilgarriff et al. 2008), l'originalité de notre approche est de prendre en considéra-

tion l'ensemble des CRC linguistiques et conceptuelles. Nous avons mis en œuvre les deux notions de contextes riches en connaissances conceptuelles (CRCC) et linguistiques (CRCL) dans une stratégie unifiée d'extraction de CRC en corpus de spécialité. Les contextes obtenus ont été évalués manuellement dans un premier temps, puis exploités dans le cadre de la traduction professionnelle. Les premières évaluations réalisées ont confirmé nos hypothèses qui postulent que les contextes riches du point de vue des linguistes peuvent être perçus comme riches par les traducteurs en complément d'autres outils d'aide à la traduction. Les résultats issus de cette thèse restent préliminaires et devront être confirmés par d'autres évaluations et expériences plus significatives.

### 9.1.2 Extraction de connaissances en corpus comparables

Après avoir identifié les CRC en corpus monolingues dans un cadre de compréhension (et de traduction), nous avons poursuivi l'exploitation des CRC mais en corpus comparables spécialisés dans une perspective de révision en traduction spécialisée. Pour cela, nous avons exploité la comparabilité de ces ressources afin d'aligner les CRC identifiés auparavant.

#### Exploitation de corpus comparables

Les corpus constituent un complément idéal des ressources linguistiques conventionnelles qui sont très souvent incomplètes ou déconstextualisées (Bowker et Barlow 2004). Ils répondent idéalement aux problématiques liées à la traduction spécialisée. En effet, ils permettent de confirmer des hypothèses de traduction en recontextualisant les termes et sont essentiels à la phase de pré-traduction : les traducteurs spécialisés se doivent d'acquérir une culture technique du domaine traité, ce qu'ils font en parcourant de nombreux textes de type explicatifs en rapport avec la thématique des textes à traduire. Les travaux de McEnery et Xiao (2007) ont montré que le recours à des corpus comparables produit des traductions de meilleure qualité en ce qui concerne la compréhension d'un domaine technique, le choix des termes et l'emploi d'expressions idiomatiques. Si les corpus comparables sont des ressources bien plus abondantes que les corpus parallèles, les lexiques bilingues extraits à partir de corpus comparables sont d'une qualité bien inférieure à ce qui peut être obtenu à partir de corpus parallèles. Cette différence de résultat dans la qualité des lexiques extraits s'explique principalement par l'absence d'éléments d'ancrage dans les corpus comparables (l'alignement préalable de paragraphes, de segment, etc. n'est pas possible avec ce type de corpus).

#### CRC bilingues

En révision, nous avons étudié la possibilité de proposer, pour un terme traduit et sa traduction attestée, des CRC alignés et de même type afin d'aider à valider le choix du traducteur. Un enjeu majeur rencontré dans

ce travail tient à l'absence de points d'ancrage conventionnels dans les corpus comparables. Pour palier cette difficulté, nous avons dégagé une stratégie basée sur des hypothèses concernant la comparabilité des corpus comparables. Pour aligner les CRC, les contextes riches en connaissances conceptuelles (CRCC) ont été écartés du fait de leur rareté en corpus de spécialité, malgré leur précision. Seuls les contextes riches en connaissances linguistiques (CRCL) ont été retenus. Notre stratégie a été de :

1. aligner les CRCL sur la base des collocatifs étant donné le terme et sa traduction attestée. Ici, notre objectif n'était pas d'identifier les meilleures traductions des collocations, mais plutôt de proposer des traductions acceptables illustrant des voisinages similaires du terme et de sa traduction. Nous avons exploité les collocations comme un premier point d'ancrage identifiant des contextes similaires dans les deux langues ;
2. filtrer les CRCL sur la base de leur longueur, la présence de pronoms au début du contexte, type du contexte (affirmative, interrogative...), ainsi que la complexité du CRC. Ces critères servent de mesures d'informativité et d'utilisabilité améliorant la qualité des contextes proposés ;
3. aligner les contextes retenus de l'étape précédente en s'appuyant sur les cognats et les termes simples traduits qui représentent un deuxième point d'ancrage affinant l'alignement des CRCL.

Cette méthodologie a été mise en œuvre dans un prototype de concordanciers bilingues dédié à la révision en traduction ou plus généralement dans l'aide à la communication interlingue en corpus comparables.

#### Apport de CRC bilingues en traduction

Les premières évaluations et expériences menées auprès de réviseurs ont montré que même si leur apport n'est pas manifeste, les CRC bilingues proposés sont une piste prometteuse dans un exercice d'aide à la révision, malgré la difficulté de la tâche.

## 9.2 Perspectives

Les travaux préliminaires de Meyer (2001) et Barrière (2004) sur les CRC ont constitué la base de cette thèse. Cependant, ces travaux fondateurs n'ont pas dépassé le stade de la preuve de concept. Dans notre travail, nous en avons étendu le champ d'application, amélioré la robustesse et porté la technologie à une échelle de prototype. Toutefois, notre contribution a des limites et il reste encore plusieurs perspectives à ce travail.

### 9.2.1 Identification de CRC

Il serait utile pour le traducteur de disposer de CRC ordonnés. Pour ce faire, il faudrait attribuer à chacun des CRC un score de confiance qui pourrait être un score de fiabilité du patron de connaissances utilisé ou de la collocation en question. Cette mesure pourrait également intégrer la densité terminologique comme une mesure de pertinence. Concernant les CRCC, nous envisageons d'élargir la liste des patrons de connaissances et d'associer à chaque patron source un patron cible équivalent explicitant la même relation sémantique. Ceci permettra de mettre en œuvre l'alignement des CRCC en corpus comparables. Par ailleurs, dans notre extraction de CRCL, les collocations contiennent également certains termes complexes ayant la même structure syntaxique. Il serait nécessaire de filtrer les termes complexes et ainsi affiner nos études des collocations en domaine de spécialité. Nous envisageons également de développer un système d'apprentissage permettant d'extraire automatiquement des CRC indépendamment du corpus utilisé.

Une piste qui nous paraît intéressante serait d'étudier les contextes qui, hormis les contextes immédiats de type distributionnel, sont utiles aux traducteurs. Il s'agit des contextes partagés par différents termes (et donc sémantiquement apparentés) ou encore les contextes faisant intervenir des termes morphologiquement liés par exemple.

### 9.2.2 Évaluation et exploitation des CRC

L'identification des CRC est une tâche orientée utilisateur dont la difficulté est d'évaluer finement l'apport des CRC dans une situation de traduction. Cependant, il est logistiquement très coûteux de réaliser des expérimentations avec des traducteurs professionnels. Par ailleurs, l'évaluation manuelle des CRC est un exercice complexe qui dépend d'une certaine expertise du domaine étudié et reste plus ou moins subjectif dans l'interprétation de certaines connaissances.

Une des solutions que nous proposons est d'intégrer les CRC dans des outils d'extraction de terminologies bilingues à partir de corpus comparables. Ces outils permettent à partir d'un terme à traduire dans une langue source d'obtenir une liste ordonnée de traduction candidates dans une langue cible. Ces traductions candidates sont obtenues en comparant le contexte traduit en langue cible du terme source avec l'ensemble des contextes des termes de la langue cible. Les traductions candidates obtenues par cette approche se présentent sous la forme d'une liste « plate » sans information contextuelle permettant d'appréhender le terme dans son contexte d'utilisation. En outre, le choix de contextes signifiants est une tâche complexe de part la quantité de contextes disponibles. Notre objectif serait d'exploiter les CRC afin de reclasser ces traductions candidates en fonction des CRC éventuellement leurs types, nombres, etc. et ainsi évaluer leur utilité.

# Bibliographie


Rodrigo arcalón martínez. *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*. PhD thesis, Université de Pompeu Fabra, 2009. (Cité page 73.)

Beryl T Sue Atkins. Bilingual dictionaries : Past, present and future. *EURALEX'96 Proceedings. Göteborg : Department of Swedish, Göteborg University*, pages 515–546, 1996. (Cité page 17.)

BT Sue Atkins et Michael Rundell. *The Oxford guide to practical lexicography*. Oxford University Press, 2008. (Cité pages 18 et 41.)

Alain Auger. Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles. *PhD thesis, Université de Neuchâtel*, 1997. (Cité pages 33 et 39.)

Alain Auger et Caroline Barrière. Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Terminology : international journal of theoretical and applied issues in specialized communication*, 14 :1–19, 2008. (Cité page 73.)

Nathalie Aussenac-Gilles et Anne Condamines. Variation syntaxique et contextuelle dans la mise au point de patrons de relations sémantiques. *Filtrage sémantique*, pages 115–149, 2009. (Cité pages 33 et 62.)

Nathalie Aussenac-Gilles et Anne Condamines. Variation and Semantic Relation Interpretation : Linguistic and Processing Issues. Dans *Terminology an Knowledge Engineering*, pages 106–122, 2012. (Cité page 33.)

Vít Baisa, Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, et Pavel Rychlý. Bilingual word sketches : the translate button. Dans Andrea Abel, Chiara Vettori, et Natascia Ralli, éditeurs, *Proceedings of the 16th EURALEX International Congress*, pages 505–513, Bolzano, Italy, 2014. (Cité page 97.)

Mona Baker. Contextualization in translator-and interpreter-mediated events. *Journal of Pragmatics*, 38 :321–337, 2006. (Cité page 13.)

Michael Barlow. Paraconc : Concordance software for multilingual parallel corpora. Dans *Proceedings of the Third International Conference on Language Resources and Evaluation. Workshop on Language Resources in Translation Work and Research*, pages 20–24, 2002. (Cité page 94.)

Caroline Barrière. Knowledge-rich contexts discovery. *Proceedings of the Seventeenth Canadian Conference on Artificial Intelligence (AI'2004)*, 2004. (Cité pages 23, 36, 37, 38, 40, 50, 62 et 124.)

Sabine Bartsch. *Structural and functional properties of collocations in English : A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag, 2004. (Cité pages 27 et 28.)

Peter Bennison et Lynne Bowker. Designing a tool for exploiting bilingual comparable corpora. Dans *LREC*, 2000. (Cité page 21.)

Godelieve Berry-Rogghe. The computation of collocations and their relevance in lexical studies. Dans A.J. Aitken, R. Bailey, et N. Hamilton-Smith, éditeurs, *The Computer and Literary Studies*, pages 103–112. Edinburgh University Press, Edinburgh, 1973. (Cité pages 49 et 64.)

Douglas Biber. *Variation across speech and writing*. Cambridge University Press, 1991. (Cité page 122.)

Li Bo, Éric Gaussier, Emmanuel Morin, et Amir Hazem. Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. Dans *TALN 2011-Conférence sur le Traitement Automatique des Langues Naturelles*, volume 1, pages 211–222, 2011. (Cité page 58.)

Claudine Bodson. *Termes et relations sémantiques en corpus spécialisés : rapport entre patrons de relations sémantiques (PRS) et types sémantiques (TS)*. Université de Montréal, 2005. (Cité page 71.)

Julien Bourdaillet, Stéphane Huet, Philippe Langlais, et Guy Lapalme. Transsearch : from a bilingual concordancer to a translation finder. *Machine Translation*, 24 :241–271, 2010. (Cité pages 94, 96 et 97.)

Didier Bourigault. *Lexter : un Logiciel d'EXtraction de TERminologie : application à l'acquisition des connaissances à partir de textes*. PhD thesis, Ecole des Hautes Études en Sciences Sociales, Paris, 1994. (Cité page 11.)

Didier Bourigault, Nathalie Aussenac-Gilles, et Jean Charlet. Construction de ressources terminologiques ou ontologiques à partir de textes Un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18 :87–110, 2004. (Cité page 21.)

Lynne Bowker. Terminology tools for translators. *BENJAMINS TRANSLATION LIBRARY*, 35 :49–66, 2003. (Cité page 21.)

Lynne Bowker. Official language minority communities, machine translation, and translator education : Reflections on the status quo and considerations for the future. *TTR : Traduction, terminologie, rédaction*, 21 : 15–61, 2008. (Cité page 12.)

Lynne Bowker. Off the record and on the fly : Examining the impact of corpora on terminographic practice in the context of translation. *Corpus-based Translation Studies : Research and Applications. London/New York : Continuum*, pages 211–236, 2011. (Cité pages 3, 14, 16, 17, 18 et 59.)

Lynne Bowker. Meeting the needs of translators in the age of e-lexicography : Exploring the possibilities. *Electronic lexicography*, pages 379–387, 2012. (Cité pages 14 et 18.)

Lynne Bowker et Michael Barlow. Bilingual concordancers and translation memories : A comparative evaluation. Dans *Proceedings of the second international workshop on language resources for translation work, research and training*, pages 70–79. Association for Computational Linguistics, 2004. (Cité pages 11, 93, 94, 95, 97 et 123.)

Lynne Bowker et Jennifer Pearson. *Working with specialized language : a practical guide to using corpora*. Routledge, 2002. (Cité page 57.)

Antoine Bride, Tim Van de Cruys, et Nicholas Asher. A generalisation of lexical functions for composition in distributional semantics. Dans *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 281–291, Beijing, China, July 2015. (Cité page 50.)

Marie Teresa Cabré. *La terminologie : théorie, méthode et applications*. Presses de l'Université d'Ottawa et Armand Colin, 1998. (Cité page 24.)

Anne Condamines. Corpus analysis and conceptual relation patterns. *Terminology*, 8 :141–162, 2002. (Cité page 71.)

Anne Condamines. Expression de la méronymie dans les petites annonces immobilières : comparaison français/anglais/espagnol. *Journal of French Language Studies*, 19 :3–23, 2009. (Cité page 62.)

Anne Condamines et Pascal Amsili. Terminology Between Language and Knowledge : An Example of Terminological Knowledge Base. Dans *Terminology and Knowledge Engineering*, pages 316–323, 1993. (Cité page 21.)

Anne Condamines et Josette Rebeyrolle. Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. *Ingénierie des Connaissances, évolutions récentes et nouveaux défis. Paris : Eyrolles*, pages 225–242, 2000. (Cité pages 62, 71 et 73.)

Anne Condamines et Josette Rebeyrolle. Searching for and identifying conceptual relationships via a corpus-based approach to a terminological knowledge base (ctkb). *Recent Advances in Computational Terminology*, pages 127–148, 2001. (Cité page 62.)

Damien Cram et Béatrice Daille. Termsuite : Terminology extraction with term variant detection. *Association for Computational Linguistics (ACL 2016)*, page 13, 2016. (Cité page 63.)

D Alan Cruse. *Lexical semantics*. Cambridge University Press, 1986. (Cité page 28.)

D Alan Cruse. Hyponymy and its varieties. Dans *The semantics of relationships*, pages 3–21. Springer, 2002. (Cité pages 61 et 73.)

Béatrice Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Université de Paris 7, Paris, 1994. (Cité page 11.)

Béatrice Daille. Extraction de collocation à partir de textes. *Traitement automatique des langues naturelles (TALN 2001)*, 2001. (Cité page 28.)

Béatrice Daille, Éric Gaussier, et Jean-Marc Langé. Towards automatic extraction of monolingual and bilingual terminology. Dans *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 515–521. Association for Computational Linguistics, 1994. (Cité page 11.)

Béatrice Daille et Emmanuel Morin. French-English terminology extraction from comparable corpora. Dans *Natural Language Processing–IJCNLP 2005*, pages 707–718. Springer, 2005. (Cité page 57.)

GILE Daniel. Les outils documentaires du traducteur. *Le traducteur et ses instruments*, 8 :73, 1993. (Cité page 58.)

Mark Davies. *The corpus of contemporary American English*. BYE, Brigham Young University, 2008. (Cité page 73.)

Hervé Déjean et Eric Gaussier. Une nouvelle approchea l'extraction de lexiques bilinguesa partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22, 2002. (Cité pages 57 et 58.)

Jean Delisle, Hannelore Lee-Jahnke, et Monique C Cormier. *Terminologie de la Traduction : Translation Terminology. Terminología de la Traducción. Terminologie der Übersetzung*, volume 1. John Benjamins Publishing, 1999. (Cité page 91.)

Estelle Delpech. Evaluation of terminologies acquired from comparable corpora : an application perspective. Dans *NODALIDA 2011*, pages 66–73, 2011. (Cité page 58.)

Estelle Delpech. *Traduction assistée par ordinateur et corpus comparables : contributions à la traduction compositionnelle*. PhD thesis, Université de Nantes, 2013. (Cité page 11.)

Alain Désilets, Christiane Melançon, Geneviève Patenaude, et Louise Brunette. How translators use tools and resources to resolve translation problems : An ethnographic study. *Proceedings of Beyond Translation Memories : New Tools for Translators MT Summit XII 2009*, 2009. (Cité page 12.)

Jörg Didakowski, Lothar Lemnitzer, et Alexander Geyken. Automatic example sentence extraction for a contemporary German dictionary. Dans *Proceedings of the 15th EURALEX International Congress*, pages 343–349, Oslo, Norway, 2012. (Cité pages 9, 42, 44, 45, 46, 48, 50, 64, 103, 107 et 122.)

Isabel Duran-Muñoz. Specialized lexicographical resources : a survey of translators' needs. *Granger, Sylviane & Magali Paquot (eds.)*, pages 55–66, 2010. (Cité pages 15 et 16.)

Christine Durieux. *Fondement didactique de la traduction technique*. Didier erudition, 1988. (Cité pages 3 et 58.)

Stefan Evert. *The Statistics of Word Co-Occurrences : Word Pairs and Collocations*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Université de Stuttgart, Stuttgart, Germany, 2005. (Cité page 50.)

Stefan Evert et Brigitte Krenn. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, 19 :450–466, 2005. (Cité page 64.)

Ismail Fahmi et Gosse Bouma. Learning to identify definitions using syntactic features. Dans *Proceedings of the EACL 2006 workshop on Learning Structured Information in Natural Language Applications*, pages 64–71, 2006. (Cité page 41.)

Robert M. Fano. *Transmission of Information : A Statistical Theory of Communication*. MIT Press, 1961. (Cité pages 49 et 64.)

Christine Fellbaum. *WordNet : An electronic lexical database*. MIT Press, 1998. (Cité pages 48 et 64.)

J R Firth. A synopsis of linguistic theory 1930-55. *The Philological Society*, 1952-59 :1–32, 1957. (Cité page 57.)

John Flowerdew. Definitions in science lectures. *Applied Linguistics*, 13(2) : 202–221, 1992a. (Cité pages 24 et 61.)

John L Flowerdew. Salience in the performance of one speech act : the case of definitions. *Discourse processes*, 15 :165–181, 1992b. (Cité pages 24 et 61.)

Bernard Fradin et Jean-Marie Marandin. Autour de la définition : de la lexicographie à la sémantique. *Langue française*, pages 60–83, 1979. (Cité page 24.)

Atsushi Fujii et Tetsuya Ishikawa. Utilizing the World Wide Web as an encyclopedia : Extracting term descriptions from semi-structured texts. Dans *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 488–495. Association for Computational Linguistics, 2000. (Cité pages 9 et 34.)

Pascale Fung et Kathleen McKeown. Finding terminology translations from non-parallel corpora. Dans *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202, 1997. (Cité pages 57 et 58.)

William A Gale et Kenneth W Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19 :75–102, 1993. (Cité pages 94 et 95.)

Daniela Garcia. *Analyse automatique des textes pour l'organisation causale des actions : réalisation du système informatique COATIS*. PhD thesis, Université de Paris 4, 1998. (Cité page 73.)

Éric Gaussier, David Hull, et Salah Aït-Mokhtar. Term alignment in use. Dans *Parallel text processing*, pages 253–274. Springer, 2000. (Cité page 11.)

Lee Gillam, Mariam Tariq, et Khurshid Ahmad. Terminology and the construction of ontology. *Terminology*, 11 :55–81, 2005. (Cité page 22.)

Daniel Gondoin. Localisation de sites Web : contraintes et enjeux. Dans Elisabeth Lavault-Olléon, éditeur, *Traduction spécialisée : pratiques, théories, formations*, pages 179–188. Peter Lang, Bern, 2007. (Cité page 91.)

Daniel Gouadec. *Profession : traducteur*. La Maison du dictionnaire, 2002. (Cité page 91.)

Gustave Guillaume. *Foundations for a Science of Language*. Amsterdam studies in the theory and history of linguistic science. Series IV, Current issues in linguistic theory. John Benjamins Publishing, Amsterdam, 1984. (Cité page 65.)

S Hagen, J Foreman-Peck, S Davila-Philipon, et B Nordgen. Elan : Effects on the european economy of shortages of foreign languages skill in entreprise. *CLIT, the National Centre for Languages*, 2006. (Cité pages 1 et 2.)

M.A.K. Halliday. *Categories of the Theory of Grammar*. The Bobbs-Merrill Reprint Series in Language and Linguistic Language. Linguistic Circle, 1961. (Cité page 49.)

Michael AK Halliday. The notion of "context" in language education. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pages 1–24, 1999. (Cité page 13.)

Michael Alexander Kirkwood Halliday. *Language as social semiotic*. London Arnold, 1978. (Cité page 122.)

F.J Hausmann. Le dictionnaire de collocations. In Hausmann FJ. , *Reichmann O. Wiegand H.E., Zgusta L.(eds),Wörterbücher*, 17 :187–195, 1989. (Cité pages 27 et 28.)

Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. Dans *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992. (Cité pages 62 et 72.)

Adrien Hermans. La définition des termes scientifiques. *Meta : Journal des traducteursMeta :Translators' Journal*, 34 :529–532, 1989. (Cité pages 24 et 61.)

Makoto Iwayama et Takenobu Tokunaga. Hierarchical bayesian clustering for automatic text classification. Dans *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, pages 1322–1327, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. (Cité page 34.)

Christian Jacquemin. A symbolic and surgical acquisition of terms through variation. Dans *International Joint Conference on Artificial Intelligence*, pages 425–438. Springer, 1995. (Cité page 11.)

Marie-Paule Jacques et Nathalie Aussenac-Gilles. Variabilité des performances des outils de tal et genre textuel. *Traitement Automatique des Langues*, 47 :11–32, 2006. (Cité page 71.)

Amélie Josselin-Leray. *Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues : étude d'un domaine de spécialité : volcanologie*. PhD thesis, Université de Lyon 2, 2005. (Cité pages 15 et 72.)

Amélie Josselin-Leray, Cécile Fabre, Josette Rebeyrolle, Aurélie Picton, et Emmanuel Planas. Good Contexts for Translators - A First Account of the Cristal Project. Dans Institute for Specialised Communication EURAC et Multilingualism, éditeurs, *Euralex 2014 : The User in Focus*, Bolzano, Italy, 2014. (Cité pages 14 et 82.)

Noriko Kando, Kazuko Kuriyama, et Toshihiko Nozue. Nacsis test collection workshop (ntcir-1)(poster abstract). Dans *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–300. ACM, 1999. (Cité page 34.)

Adam Kilgarriff. Which words are particularly characteristic of a text ? a survey of statistical approaches. *Language Engineering for Document Analysis and Recognition*, pages 33–40, 1996. (Cité page 50.)

Adam Kilgarriff, Pavel Rychlý, Miloš Husák, Michael Rundell, et Katy McAdam. GDEX : Automatically finding good dictionary examples in a corpus. Dans *Proceedings of the XIII EURALEX International Congress*, pages 425–432, Barcelona, 2008. (Cité pages 9, 41, 42, 43, 44, 46, 48, 50, 64, 86, 103, 107 et 122.)

Richard Kittredge. Variation and homogeneity of sublanguages. *Sublanguage : studies of language in restricted semantic domains*, pages 107–137, 1982. (Cité page 122.)

Alexander Künzli. Experts versus novices : l'utilisation de sources d'information pendant le processus de traduction. *Meta : Journal des traducteursMeta :Translators' Journal*, 46 :507–523, 2001. (Cité page 12.)

Louise Larivière. Comment formuler une définition terminologique. *Meta*, 41(3) :405–418, 1996. (Cité page 24.)

Audrey Laroche et Philippe Langlais. Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. Dans Chu-Ren Huang et Dan Jurafsky, éditeurs, *COLING*, pages 617–625. Tsinghua University Press, 2010. (Cité page 57.)

Ora Lassila et Ralph R. Swick. Resource Description Framework(RDF) Model and Syntax Specification, 1999. (Cité page 60.)

Luce Lefeuvre et Anne Condamines. Constitution d'une base bilingue de marqueurs de relations concep-tuelles pour l'élaboration de ressources termino-ontologiques. *Proceedings of the conference Terminology and Artificial Intelligence*, 2015. (Cité pages 72, 73 et 122.)

Stéphanie Léon. *Acquisition automatique de traductions d'unités lexicales complexes à partir du Web*. PhD thesis, Université de Provence, Aix-Marseille I, 2008. (Cité page 104.)

Marie-Claude L'Homme. Le dicoinfo. méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire*, 217 :78–103, 2008. (Cité page 26.)

Véronique Malaisé, Pierre Zweigenbaum, et Bruno Bachimont. Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. *Actes de TALN*, pages 269–278, 2004. (Cité pages 39 et 75.)

L Mammino. Il linguaggio e la scienza. *Torino : Societá Editrice Internazionale*, 1995. (Cité page 100.)

Elizabeth Marshman. *Lexical Knowledge Patterns for Semi-automatic Extraction of Cause-effect and Association Relations from Medical Texts : A Comparative Study of English and French*. Université de Montréal, 2007. (Cité page 71.)

Elizabeth Marshman. Enriching terminology resources with knowledge-rich contexts : A case study. *Terminology*, 20 :225–249, 2014. (Cité pages 60, 86, 107 et 122.)

Elizabeth Marshman, Julie L Gariépy, et Charissa Harms. Helping language professionals relate to terms : Terminological relations and termbases. *JoSTrans*, 18, 2012. (Cité page 14.)

Elizabeth Marshman, Marie-Claude L'Homme, et Victoria Surtees. Portability of cause-effect relation markers across specialised domains and text genres : a comparative evaluation. *Corpora*, 3 :141–172, 2008. (Cité page 61.)

R.A. Martínez, G.S. Martínez, Aplicada. Institut Universitari de Lingüística, et C. Bach. *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*. PhD Thesis. Université de Pompeu Fabra, 2009. (Cité page 64.)

A McEnery et Richard Xiao. Parallel and comparable corpora : What is happening. *Incorporating Corpora. The Linguist and the Translator*, pages 18–31, 2007. (Cité page 123.)

Kathleen McKeown et Dragomir Radev. Collocations. *Handbook of Natural Language Processing. Marcel Dekker*, 2000. (Cité page 50.)

I Dan Melamed. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25 :107–130, 1999. (Cité page 11.)

Alan K Melby, Christopher Foster, et al. Context in translation : Definition, access and teamwork. *Translation & Interpreting*, 2 :1–15, 2010. (Cité page 13.)

Igor Mel'ćuk. Collocations and lexical functions. *2001 [1998]*, pages 23–54, 1998. (Cité pages 27 et 28.)

Ingrid Meyer. Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. Dans Didier Bourigault, Christian Jacquemin, et Marie-Claude L'Homme, éditeurs, *Recent Advances in Computational Terminology*, pages 279–302. John Benjamins Publishing Company, 2001. (Cité pages 4, 21, 22, 23, 24, 26, 33, 62, 64, 75, 121 et 124.)

Ingrid Meyer, Douglas Skuce, Lynne Bowker, et Karen Eck. Towards a new generation of terminological resources : an experiment in building a terminological knowledge base. Dans *Proceedings of the 14th conference on Computational linguistics-Volume 3*, pages 956–960. Association for Computational Linguistics, 1992. (Cité page 21.)

Emmanuel Morin. Des patrons lexico-syntaxiques pour aider au dépouillement terminologique. *Traitement Automatique des Langues*, 40 : 143–166, 1999. (Cité page 75.)

Emmanuel Morin et Béatrice Daille. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44 :79–95, 2009. (Cité page 102.)

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, et Kyo Kageura. Bilingual terminology mining-using brain, not brawn comparable corpora. Dans *Annual Meeting-Association for Computational Linguistics*, volume 45, page 664, 2007. (Cité page 58.)

Katell Morin-Hernandez. *Revision as a key function of translation quality management in a professional context*. Theses, Université Rennes 2, Janvier 2009. (Cité page 92.)

Isabel Durán Munoz. Meeting translators' needs : translation-oriented terminological management and applications. *The Journal of Specialised Translation*, 18 :77–92, 2012. (Cité page 16.)

Smaranda Muresan et Judith Klavans. A Method for Automatically Building and Evaluating Dictionary Resources. Dans *LREC*. European Language Resources Association, 2002. (Cité page 33.)

Maria Teresa Musacchio et Giuseppe Palumbo. Shades of Grey : A Corpus-driven Analysis of LSP Phraseology for Translation Purposes. Dans C. Taylor, K. Ackerley, et E. Castello, éditeurs, *Corpora for University Language Teachers*, pages 69–79. Peter Lang, Bern, 2008. (Cité page 101.)

Makato Nagao. Jiten Kêshiki deno senmon bunya no chishiki no taikêteki kôbunhô (Organisation systématique du savoir d'un domaine spécifique exprimé sous la forme dictionnairique). *Jinkôchinô gakkaishi (Revue de la Société d'intelligence artificielle)*, 7 :320–328, 1992. (Cité page 25.)

Roberto Navigli et Paola Velardi. Learning Word-Class Lattices for Definition and Hypernym Extraction. Dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1327, Uppsala, Sweden, July 2010. Association for Computational Linguistics. (Cité pages 9, 46, 47 et 50.)

Luka Nerima, Violeta Seretan, et Eric Wehrli. Le problème des collocations en tal. *Cahiers de linguistique française*, 27 :65–115, 2006. (Cité pages 48, 49 et 50.)

Peter Newmark. *A Textbook of Translation*. Prentice-Hall International, 1988. (Cité page 101.)

M. Noailly. *L'adjectif en français*. Collection L'essentiel français. Ophrys, 1999. (Cité page 65.)

Patrick Pantel et Marco Pennacchiotti. Automatically harvesting and ontologizing semantic relations. *Ontology learning and population : Bridging the gap between text and knowledge*, pages 171–198, 2008. (Cité page 74.)

Elsa Pascual et Marie-Paule Péry-Woodley. La définition dans le texte. *Textes de type consigne–Perception, action, cognition*, pages 65–88, 1995. (Cité page 25.)

Silvia Pavel. Neology and phraseology as terminology-in-the-making. *Terminology : Applications in interdisciplinary communication*, 21 :34, 1993. (Cité page 26.)

Darren Pearce. A Comparative Evaluation of Collocation Extraction Techniques. Dans *proceedings of the Third International Conference on Language Resources and Evaluation, LREC*, 2002. (Cité page 50.)

Jennifer Pearson. *Terms in context*, volume 1. John Benjamins Publishing Compagny, Amsterdam/Philadelphia, 1998. (Cité pages 23, 24 et 61.)

Jennifer Pearson. Une tentative d'exploitation bi-directionnelle d'un corpus bilingue. *Cahiers de grammaire*, 25 :53–69, 2000. (Cité page 71.)

Jacques Permentiers, Franco Troiano, et Erik Springael. *Traduzione, adattamento ED Editing multilingue*. TCG edition Bruxel, 1996. (Cité page 101.)

Emmanuel Planas, Aurélie Picton, et Amélie Josselin-Leray. Exploring the Use and Usefulness of KRCs in Translation : Towards a Protocol. Dans *Terminology and Knowledge Engineering 2014*, page 10 p, Berlin, Germany, Juin 2014. (Cité page 82.)

Reinhard Rapp. Automatic identification of word translations from unrelated English and German corpora. Dans *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics, 1999. (Cité pages 57 et 58.)

Josette Rebeyrolle. *Forme et fonction de la définition en discours*. PhD thesis, Université de Toulouse 2, 2000a. (Cité pages 33, 39 et 40.)

Josette Rebeyrolle. Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. *Actes Journées Francophones d'Ingénierie de la Connaissance (IC'2000)*, pages 105–114, 2000b. (Cité page 24.)

Josette Rebeyrolle et Ludovic Tanguy. Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de grammaire*, 25 :153–174, 2000. (Cité pages 24 et 71.)

Melanie Reiplinger, Ulrich Schäfer, et Magdalena Wolska. Extracting Glossary Sentences from Scholarly Articles : A Comparative Evaluation of Pattern Bootstrapping and Deep Analysis. Dans *Proceedings of the Association for Computational Linguistics-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 55–65, Stroudsburg, PA, USA, 2012. (Cité page 48.)

Isabelle S Robert. *La révision en traduction : les procédures de révision et leur impact sur le produit et le processus de révision*. PhD thesis, Université d'Antwerp, 2012. (Cité page 91.)

Robert, Isabelle. Translation revision procedures : An explorative study. *Boulogne, Pieter (ed.)*, 2008. (Cité page 91.)

Roda P Roberts. Bilingual dictionaries prepared in terms of translators' needs. Dans *Proceedings of CTIC 3rd Conference, Translation in the Global Village*, pages 51–65, 1994. (Cité pages 14 et 15.)

Roda P Roberts et Jacqueline Bossé-Andrieu. Corpora and translation. *Lexicography, Terminology, and Translation–Text-based Studies in Honour of Ingrid Meyer. Ottawa : Presses de l'Université d'Ottawa*, pages 201–214, 2006. (Cité page 14.)

Michel Rochard. Traduction professionnelle et traduction pédagogique. le lien de l'enquête. *Les Cahiers de l'ILCE*, pages 126–127, 1999. (Cité page 12.)

Mathieu Roche. *Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes*. PhD thesis, Université de Paris 11, 2004. (Cité page 65.)

Jérôme Rocheteau et Béatrice Daille. Ttc termsuite : A uima application for multilingual terminology extraction from comparable corpora. Dans *5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 9–12, 2011. (Cité page 63.)

Margaret Rogers et Khurshid Ahmad. The translator and the dictionary : Beyond words. *Atkins, BT Sue (ed.)*, pages 193–204, 1998. (Cité pages 13, 14 et 15.)

Michael Rundell et Adam Kilgarriff. Automating the Creation of Dictionaries : Where Will It All End ?. Dans *A Taste for Corpora*, Studies in Corpus Linguistics (StCoL) : 45, pages 257–281. John Benjamins, 2011. (Cité page 122.)

Juan C Sager. *A practical course in terminology processing*. John Benjamins Publishing, 1990. (Cité pages 21, 24 et 25.)

Horacio Saggion. Identifying Definitions in Text Collections for Question Answering. Dans *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1927–1930, 2004. (Cité pages 9, 34, 35, 36, 40, 48 et 50.)

Anne-Kathrin Schumann. A Case Study of Knowledge-Rich Context Extraction in Russian. Dans *Proc. of the 9th International Conference on Terminology and Artificial Intelligence*, pages 142–145, Paris, France, 2011. (Cité page 107.)

Anne-Kathrin Schumann. Knowledge-rich context extraction and ranking with knowpipe. Dans Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, et Stelios Piperidis, éditeurs, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3626–3630. ELRA, 2012a. (Cité page 107.)

Anne-Kathrin Schumann. Towards the Automated Enrichment of Multilingual Terminology Databases with Knowledge-Rich Contexts–Experiments with Russian EuroTermBank Data. Dans *Proceedings of the 2nd Workshop on the Creation, Harmonization and Application of Terminology Resources (CHAT'12)*, pages 27–34, 2012b. (Cité pages 60 et 107.)

Patrick Séguéla. *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. PhD thesis, Université de Toulouse 3, 2001. (Cité pages 63 et 73.)

Violeta Seretan. *Syntax-based collocation extraction*, volume 44. Springer Science & Business Media, 2011. (Cité page 48.)

Violeta Seretan et al. *Collocation extraction based on syntactic parsing*. PhD thesis, Université de Genève, 2008. (Cité pages 48 et 50.)

Serge Sharoff. Creating general-purpose corpora using automated search engine queries. *WaCky ! Working papers on the Web as Corpus. Gedit*, pages 63–98, 2006. (Cité page 102.)

J. M. Sinclair, S. Jones, et R. Daley. *English Lexical Studies. Final Report of O.S.T.I. Programme C/LP/o8.* Department of English, 1970. (Cité pages 27 et 29.)

Angelika Storrer et Sandra Wellinghoff. Automated detection and annotation of term definitions in German text corpora. Dans *Proceedings of LREC*, volume 2006, 2006. (Cité page 46.)

Gerhard Strauss, Ulrike Hass-Zumkehr, et Gisela Harras. *Brisante W"orter von Agitation bis Zeitgeist : ein Lexikon zum "offentlichen Sprachgebrauch*. deGruyter, Berlin/New York, 1989. (Cité pages 42 et 45.)

Rita Temmerman. *Towards new ways of terminology description : the sociocognitive-approach*, volume 3. John Benjamins Publishing, 2000. (Cité page 21.)

Louis Trimble. *English for science and technology : A discourse approach*. Cambridge University Press Cambridge, 1985. (Cité page 25.)

Agnès Tutin. For an extended definition of lexical collocations. Dans *Proceedings of Euralex*, 2008. (Cité page 28.)

Teun Adrianus Van Dijk, Walter Kintsch, et Teun Adrianus Van Dijk. *Strategies of discourse comprehension*. Citeseer, 1983. (Cité page 58.)

Krista Varantola. Translators and their use of dictionaries. *BTS Atkins Using Dictionaries. Tübingen, Niemeyer*, pages 179–192, 1998. (Cité pages 12, 14, 16 et 59.)

Krista Varantola. The contextual turn in learning to translate. *Lexicography, Terminology, and Translation : Text-based Studies in Honour of Ingrid Meyer*, page 215, 2006. (Cité pages 3 et 17.)

Jean Véronis. From the rosetta stone to the information society. Dans *Parallel Text Processing*, pages 1–24. Springer, 2000. (Cité page 11.)

Eric Wehrli. Parsing and collocations. Dans *International Conference on Natural Language Processing*, pages 272–282. Springer, 2000. (Cité page 50.)

Eline Westerhout. Extraction of Definitions Using Grammar-Enhanced Machine Learning. Dans Alex Lascarides, Claire Gardent, et Joakim Nivre, éditeurs, *EACL (Student Research Workshop)*, pages 88–96. The Association for Computer Linguistics, 2009. (Cité page 41.)

Geoffrey Williams. Collocational networks : Interlocking patterns of lexis in a corpusof plant biology research articles. *International Journal of Corpus Linguistics*, 3 :151–171, 1998. (Cité pages 22 et 26.)

Geoffrey Williams. Traduction et corpus, corpus et recherche. *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut*, 27 :69–79, 2008. (Cité page 4.)

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

# Thèse de Doctorat

Firas HMIDA

## Identification et exploitation de contextes riches en connaissances pour l'aide à la traduction terminologique

## Identification and exploitation of knowledge-rich contexts for terminological assisted translation


### Résumé

Les outils de traduction assistée par ordinateur et de gestion terminologique sont le plus souvent utilisés pour répondre au besoin de la gestion de l'écrit multilingue et monolingue. En effet, ils facilitent l'accès aux termes techniques et aux expressions liés à des domaines de spécialité, et indispensables à tout processus de communication. La compréhension de ces expressions techniques peut être potentialisée au moyen de leur « contextualisation ». Néanmoins, avoir accès à un terme ou à sa traduction ne suffit pas, encore faut-il être capable de l'employer correctement et d'en appréhender le sens exact. Cette contextualisation a donc lieu à deux niveaux : dans les textes et dans la terminologie. Au niveau textuel, l'utilisateur doit avoir accès à des informations concernant l'usage des termes, à savoir des contextes riches en connaissances linguistiques. Au niveau terminologique, il doit avoir accès aux relations sémantiques ou conceptuelles entre termes afin de mieux en saisir le sens, à savoir des contextes riches en connaissances conceptuelles. Dans le cadre de cette thèse, nous avons proposé une stratégie d'extraction de contextes riches en connaissances (CRC) permettant de produire un premier prototype de dictionnaires terminologiques. Nous avons poursuivi nos travaux dans un cadre bilingue et plus particulièrement en phase de révision du processus de traduction spécialisée. Nous avons proposé une méthodologie d'élaboration d'un concordancier bilingue fournissant des CRC alignés à partir de corpus comparables spécialisés. Les évaluations menées montrent que les CRC proposés sont utiles malgré la difficulté de l'exercice étudié.



### Mots clés

Traduction assistée par ordinateur, Traduction spécialisée, Contexte riche en connaissance, corpus comparable spécialisés, Terminologie.



### Abstract

Computer-assisted translation and terminology management tools are often used to meet the needs in management of multilingual and monolingual writings. These tools facilitate the access to technical terms and expressions that are related to areas of specialty, and essential to any communication process. The understanding of technical terms can be potentiated by their "contextualization". However, having access to a term or its translation is not enough, since it is also necessary to be able to use it properly and to understand its exact meaning. Thus, this contextualization is estabilished on two levels: in texts and in the terminology. In texts, the user must have access to information regarding the use of terms, namely linguistic knowledge-rich contexts. In the terminology, the user requires access to semantic or conceptual relationships between the terms to better understand its meaning, namely conceptual rich-knowlegde contexts. In the framework of this thesis, we proposed a strategy for extracting Knowledge-Rich Contexts (KRCs) to produce a new terminological dictionary. It is to provide, for each term and its possible translations, the KRCs in which it occurs. We continued our work in a bilingual phase part of specialized translation, under continuous revision. We propose a new generation of bilingual concordancers that take as input a term and its translation, and provides not parallel, but aligned Knowledge-Rich Contexts from specialized comparable corpora. The evaluation show that our concordancer can assist revisers despite the difficulty of the task.



### Key Words

Computer-Assisted Translation, Specialized Translation, Knowledge-Rich Context, Specialized Comparable Corpora, Terminology.