Évaluation de solutions de traduction pour les services Semantia

Lucile Paroz

▶ **To cite this version:**

Lucile Paroz. Évaluation de solutions de traduction pour les services Semantia. Linguistique. 2010. dumas-00677687

HAL Id: dumas-00677687

https://dumas.ccsd.cnrs.fr/dumas-00677687v1

Submitted on 9 Mar 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Stendhal – Grenoble 3
1180, avenue Centrale
38400 Saint Martin d'Hères

Parc d'Activités de Gémenos
30, avenue du Château de Jouques
Les Espaces de la Ste Baume
13420 Gémenos

Évaluation de solutions de traduction pour les services Semantia

PAROZ Lucile

Mémoire de Master 2 Professionnel Industries de la Langue

Parcours Traitement Automatique du Langage Écrit et Parlé

UFR Sciences du Langage - DIP

Sous la direction de M. Georges ANTONIADIS

Année universitaire 2009-2010

MOTS-CLÉS : traduction, traduction automatique, base de connaissances

RÉSUMÉ

Le domaine de la traduction automatique ne propose pas d'outils parfaits, mais ils se révèlent utiles pour saisir le sens général d'un texte. Ils sont généralement plus performants lorsque les textes traitent d'un domaine en particulier.

Les bases de connaissances de Semantia sont justement crées pour un domaine prédéfini. Nous nous sommes interrogés sur la façon d'adapter une base de connaissances Semantia à d'autres langues.

Deux solutions ont été envisagées et vont être évaluées. Nous pouvons disposer de plusieurs bases de connaissances et à chacune d'entre elles correspondra une langue, ou nous pouvons utiliser un système de traduction automatique avec une seule base de connaissances.

KEYWORDS : translation, machine translation, knowledge base

ABSTRACT

The domain of the machine translation does not offer perfect tools, but they can be useful to understand the general sense of a text. They are generally more successful when texts deal with a specific domain.

The knowledge bases of Semantia are designed for a predefined domain. We are wondering how to adapt a knowledge base to the other languages.

Two solutions were envisaged and are going to be tested. We can arrange several knowledge bases and each of them will correspond to a language, or we can use a machine translation system with a single knowledge base.

Quelques traductions vues sur Internet

dessiccateur de dégringolade

Traduction automatique de *tumble dryer* (sèche-linge),

Systran

services de massage effectué par lettre recommandée

Traduction automatique de *[…] selection of massage services performed by registered, in-house massage therapists*,

Google Traduction

Note de confidentialité

Certaines données et informations de ce mémoire, qu'elles soient explicites, sous-entendues ou masquées, sont strictement confidentielles. Il s'agit de données internes à l'entreprise ayant accueilli le stagiaire.

Afin de respecter le travail du personnel et la stratégie de l'entreprise, nous vous prions de ne diffuser celles-ci en aucun cas ; sauf accord préalable de la direction générale de Semantia.

Remerciements

Je tiens à remercier les personnes qui m'ont épaulé avant et pendant ce stage :

- M. Jean-Jacques Schneider, Mme Sophie Girot et toute l'équipe de Semantia qui ont largement contribué à rendre ma vie en entreprise agréable et motivante ;
- Mon responsable de stage à l'Université, M. Georges Antoniadis, pour sa patience et sa disponibilité ;
- et M. Jean Véronis, qui fut d'une aide précieuse dans la recherche d'un stage de fin d'études ;

Merci également à tous les étudiants talistes et taliens que j'ai croisé sur mon chemin universitaire, je pense tout particulièrement à Mathieu Allione, Charlotte Danesi, Emilie Nougué, Marlène Omani, Manon Quintana, Laurie Serrano et Laure Voelckel.

Enfin, je remercie ma famille pour tout le soutien qu'elle m'a apporté, pendant ces cinq années d'études après le baccalauréat.

Sommaire

Introduction

Si ce mémoire commence par une note d'humour sur la traduction automatique, elle n'en reste pas moins un des domaines du Traitement Automatique du Langage Naturel (ou TALN) des plus difficiles à appréhender. La lisibilité d'une traduction effectuée par un ordinateur est souvent réduite et le panel des langues disponibles est plutôt modeste. La complexité du langage humain demeure inaccessible, malgré plus d'un demi-siècle de recherches en traduction automatique.

Semantia est une société en demande d'élargissements linguistiques ; son objectif est l'amélioration des services proposés aux clients. Pour disposer de services multilingues, la question de la pertinence de la traduction humaine par rapport à une traduction automatique est primordiale. Au cours du stage que nous avons effectué, nous avons été confronté aux problèmes que soulève la traduction. Ils se sont posés dans une moindre mesure puisque notre sujet concerne un domaine - et donc un vocabulaire – précis. De plus, le champ de recherches est réduit aux seules technologies Semantia, en l'occurrence des bases de connaissances.

Ce projet présente une évaluation de deux solutions de traduction, la première nécessitant l'existence d'une base de connaissances par langue (traduction humaine) et la seconde faisant appel à la traduction automatique pour la création d'une base de connaissances multilingue.

Partie 1

Contexte et problématique

Chapitre 1 – Présentation de l'entreprise

Semantia est une société développant des technologies de Traitement Automatique du Langage - dorénavant, TAL - et proposant ses services aux entreprises.

A - Historique

Implantée dans le Parc d'Activités de Gémenos, près de Marseille, cette entreprise a été créée en juillet 2000. Elle dispose aussi d'une agence sur Paris, près des Champs-Élysées.

L'équipe fondatrice de Semantia possédait des connaissances dans le domaine du Traitement Automatique du Langage Naturel (ou TALN), principalement, dans le dialogue Homme/Machine. Les trois premières années ont été consacrées au développement et à l'amélioration d'une technologie originale de gestion et d'anticipation de dialogue Homme/Machine.

Fin 2003, une fois ce noyau technologique mis en place, Semantia put créer des services autour de celui-ci avec l'extraction de concepts et de critères dans des courriers électroniques, des petites annonces ou bien des fiches produits. Ils sont alors développés chez les premiers clients. Leurs données sont organisées et accessibles de manière pertinente et intelligente, quelque soit le domaine ; les services sont multi-canaux (Web Agent, moteur de recherche...) et multi-sources (courriers électroniques, papiers ou formulaires). Le premier client à en bénéficier est la SNCM en 2001; deux ans plus tard, EDF et GDF deviennent les premières références nationales pour Semantia. Avec une moyenne de 6500 dialogues par jour depuis 2007, EDF est un client important en terme de volumétrie; tout comme SPIR qui traite 50000 petites annonces par jour grâce à cette technologie.

De part son positionnement large (multi-canaux, multi-sources et multi-domaines), Semantia est au cœur de la relation client, l'entreprise dispose d'une offre de solutions complètes, clés en main. Elle devient le maillon indispensable de la communication entreprises/consommateurs en proposant un outil relationnel intelligent et performant, centré sur la compréhension du besoin des clients.

B - Prestations de services, clients

Comme dit précédemment, les services Semantia permettent d'extraire du sens des données en les organisant de manière pertinente.

Nous distinguons cinq types de services complémentaires :

- Indexation intelligente;

L'indexation intelligente soumet les données de la société cliente à des référentiels complets et ciblés sur un domaine précis, *automobile* ou *téléphonie* pour *Topannonces.fr,* par exemple.

- Moteur de recherche intelligent;

Interfacée avec un moteur de recherche, cette indexation permet aux internautes de disposer d'un outil d'interrogation des données.

- Modération automatique;

Pour que les données traitées soient cohérentes et de qualité, le service de modération, toujours automatisé, s'appuie sur un cahier juridique pour détecter les informations interdites, hors sujet ou tout simplement non pertinentes. Il est alors possible de contrôler, modérer ou corriger leur diffusion.

- Traitement des e-mails;

Le traitement automatique des courriers électroniques permet de répondre aux demandes faites à la société cliente. La réaction des services est immédiate et pertinente, elle se repose sur un répertoire (ou book) de réponses. Ce fonctionnement permet d'assurer la pérennité des réponses proposées.

- Web agent.

Le Web Agent permet à l'internaute de dialoguer en langage naturel; cet interlocuteur, certes virtuel, est spécialiste du domaine concerné : il aide à la navigation sur le site internet de la société cliente et répond aux demandes d'assistance. Ce support est innovant et instaure un dialogue interactif avec l'internaute sans besoin d'intervention humaine. Offrant des réponses pertinentes, le Web Agent est un bon outil de fidélisation de la clientèle.

Toutes ces technologies sont articulées autour de bases de connaissances (nous verrons au chapitre 3 de cette partie comment elles s'organisent) qui décrivent le domaine particulier du client. Les sociétés travaillant avec Semantia se répartissent sur des secteurs d'activités variés, un grand nombre d'entreprises est concerné par les besoins auxquels répondent ses services.

Par exemple, les domaines de la **télécommunication** sont très touchés par les phénomènes de mise en défaut de leurs services d'assistance, notamment les fournisseurs d'accès Internet et de téléphonie - comme Neuf Télécom -, Semantia leur permet d'améliorer et d'étendre cette relation client.

De même, les sociétés de **vente à distance** (en ligne ou par catalogue) se situent dans un marché fortement concurrentiel où la simple exposition de produits ne suffit plus, ce sont les services proposés autour de la vente qui vont fidéliser les consommateurs.

La **presse** et les **médias -** avec des clients comme SPIR – y trouvent aussi leur compte puisqu'ils manipulent d'immenses volumes de données.

Semantia est aussi présente sur le **marché de l'énergie** avec EDF, GrDF et GDF-SUEZ où un Web agent permet une relation client plus suivie.

D'autres domaines sont concernés par ses services comme le **secteur bancaire** ou les **assurances**...

Fig. 1 : Capture d'écran de Laura, conseillère virtuelle EDF utilisant la technologie Semantia

C - Marché et concurrence

Aujourd'hui, Semantia est la seule du marché à proposer une couverture complète de la compréhension des besoins clients : multi-canaux, multi-sources et multi-domaines.

Quelques entreprises proposent des services proches, comme Conviviance, Artificial Solutions, Viavoo, Cantoche, ou bien Virtuoz...

Viavoo, par exemple, se positionne sur le marché de l'aide à la relation client. Cette entreprise parisienne propose deux solutions additionnelles :

- un outil d'analyse des remontées clients qu'elles soient internes (formulaires web...) ou externes (réseaux sociaux...) à la société cliente,
- et un outil collaboratif de gestion de ces remontées clients, non automatisé.

Si ces solutions ont l'avantage de collecter des informations depuis de multiples sources, elles nécessitent une intervention humaine régulière et donc coûteuse. Par conséquent, le dialogue direct avec les consommateurs n'est pas systématique : Viavoo fournit une analyse par verbatim de ce qu'elle comprend, mais c'est à la société utilisatrice de s'occuper des retours à faire à ses clients.

Virtuoz est une entreprise qui crée des agents virtuels ; elle dispose d'une technologie de type dialogue Homme/Machine ainsi que de l'interface utilisateur correspondante. Celle-ci peut prendre la forme d'un avatar, d'un panneau de recherche, d'une fenêtre de chat ou d'une foire aux questions (FAQ), tout comme les services Semantia. Basé sur des graphes conceptuels [Sowa, J.F., 1976, 336–357], le système, certes performant, nécessite malgré tout une forte implication des sociétés clientes.

Face à une concurrence morcelée, Semantia propose une offre complète qui permet aux entreprises de pérenniser leurs informations et de fidéliser leur clientèle.

D - Organisation générale

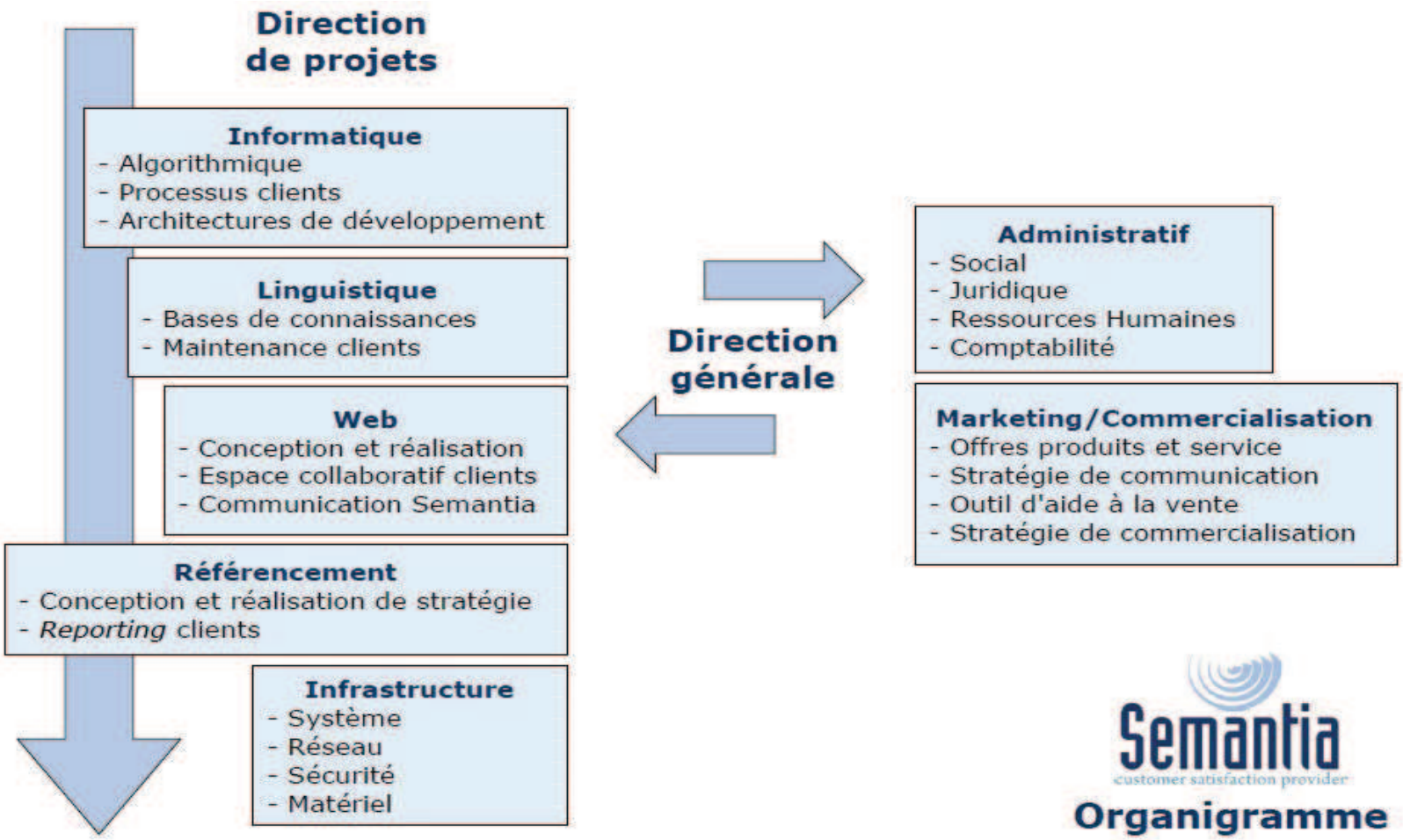

Fig. 2 : Organigramme Semantia

Chapitre 2 – Objet d'étude

Pour rester performante, Semantia travaille sans cesse à l'amélioration des services qu'elle propose. En particulier, il est indispensable qu'ils soient adaptables à d'autres langues pour élargir le marché de Semantia. Le stage effectué rejoint ce besoin, nous allons décrire et expliciter notre objet d'étude à travers le sujet de stage et la problématique du projet.

A - Sujet de stage

L'objectif est de mettre en place une architecture permettant de tester et de valider une solution de traduction en différentes langues pour les services Semantia. Nous devrons faire le choix entre deux solutions possibles :

- Intégrer dans la partie linguistique du moteur Semantia les règles nécessaires à la prise en compte des différentes langues,
- Mettre en amont du service Semantia un module de traduction automatique basé sur des solutions du commerce, et développer les règles nécessaires.

Pour la première solution, nous définirons avec l'équipe linguistique une méthodologie permettant de minimiser la complexité des règles spécifiques pour chaque langue.

Pour la seconde solution, nous identifierons les règles spécifiques à mettre en place pour corriger les erreurs des traductions.

En collaboration avec la direction, une grille d'évaluation des deux méthodes sera mise en place en tenant compte des impacts humains et financiers de chacune d'entre elles.

Dans le cadre du stage seules deux langues seront étudiées, mais le principe retenu pourra être étendu à au moins cinq langues.

1) Quelle approche ?

Les services Semantia fonctionnent autour d'un moteur s'appuyant sur une base de connaissances. De part ce fonctionnement, les solutions de traduction que nous allons mettre en place concerneront uniquement les bases de connaissances, qui sont en français. Nous travaillerons à leur niveau car ce sont elles qui gèrent la partie linguistique du moteur Semantia. L'architecture créée permettra la comparaison et l'évaluation des deux solutions possibles pour la traduction.

En ce qui concerne la première, il s'agira de créer une base de connaissances appropriée pour chaque nouvelle langue, la base de connaissance française concernée sera dupliquée dans d'autres langues. Semantia aurait donc à disposition une solution multilingue sous la forme de plusieurs bases monolingues.

La seconde solution nécessite l'utilisation d'un service de traduction automatique venant du commerce. Amélioré grâce à quelques règles, ce module serait appelé lors d'un passage dans une base de connaissance monolingue. La traduction automatique sera ajustée dans chaque langue par des règles correctrices relevant les principales erreurs observées. Dans cet optique, Semantia combinerait un module de traduction automatique à une base de connaissance monolingue.

Pour développer l'objet d'étude de manière plus concrète, il faut savoir que nous allons nous appuyer sur une base de connaissance française préexistante dans le domaine de l'hôtellerie. A partir de celle-ci, nous créerons une nouvelle base de connaissance en anglais pour la méthode qui consiste à avoir un système par langue. Pour la seconde solution, nous utiliserons directement cette base de connaissance française pour y greffer un lexique de correction des erreurs de traduction automatique les plus fréquentes en anglais.

L'architecture que nous établirons sera exploitée uniquement sur le français et l'anglais, ce qui permettra de tester le temps de conception, de recherche, la difficulté éventuelle de certaines tâches...

L'évaluation mise en place prendra en compte les aspects financiers mais aussi humains, il est nécessaire que chaque impact des deux approches soit explicité pour aider à la prise de décision en ce qui concerne les solutions de traduction.

B - Hypothèses et problématique

Avant de débuter le projet, nous pouvons évoquer les questions qui se posent sur chacune des méthodes.

Tout d'abord, pour la première solution de traduction, la durée de la création et de la mise en place d'une nouvelle base de connaissances et dans une autre langue ne sera-t-elle pas trop longue et fastidieuse par rapport à un « simple » ajout de règles comme prévu pour la seconde solution ?

Néanmoins, dans cette seconde solution, l'ajout de règles à une base de connaissance pour de la traduction automatique ne risque-t-il pas de se révéler une tâche trop ardue ? Ces règles seront-t-elles trop nombreuses, pouvons-nous être dépassés par le nombre d'erreurs faites par un système de traduction automatique, malgré le domaine restreint choisi ? De plus, la difficulté de conception d'une méthode utilisant de la traduction automatique peut énormément dépendre du système que nous choisirons d'utiliser.

Nous allons tenter d'éclairer ces questions de réponses objectives. Notons que l'utilisation d'un tel lexique de règles ne devra pas entraîner une baisse dans le taux de détection de thèmes de la base de connaissances, il s'agira de veiller à ce que les deux solutions obtiennent le même niveau de compréhension.

Aux vues de ces réflexions, nous allons tenter de répondre aux questions suivantes: *Laquelle des deux solutions proposées sera la plus pertinente pour permettre aux services Semantia d'être multilingues ? Dans quelle mesure peut-on les évaluer ?*

Dans quelle mesure la pertinence d'une solution automatique de traduction sera suffisante, par rapport à un traducteur humain, pour permettre aux services Semantia d'être multilingues ?

Chapitre 3 – Technologie Semantia

Pour pouvoir détailler les étapes de la mise en place de l'architecture nécessaire à l'évaluation des solutions de traduction, il nous faut expliciter les termes spécifiques aux technologies Semantia et décrire le fonctionnement de certains de ses services.

En exposant le sujet du stage, nous évoquions la traduction des bases de connaissances, nous allons les définir et découvrir comment elles fonctionnent.

A - Le cœur de la technologie: la base de connaissances

1) Définition générale

La base de connaissances est le noyau de la technologie Semantia, c'est elle qui décrit la manière dont vont réagir les services lors de l'analyse des demandes des utilisateurs. Elle permet de mettre en relation des **locutions**, un **thème** et des **réponses**. Au lieu d'obtenir un thème avec une simple reconnaissance par mot-clé, la détection peut être élargie grâce à une série de locutions.

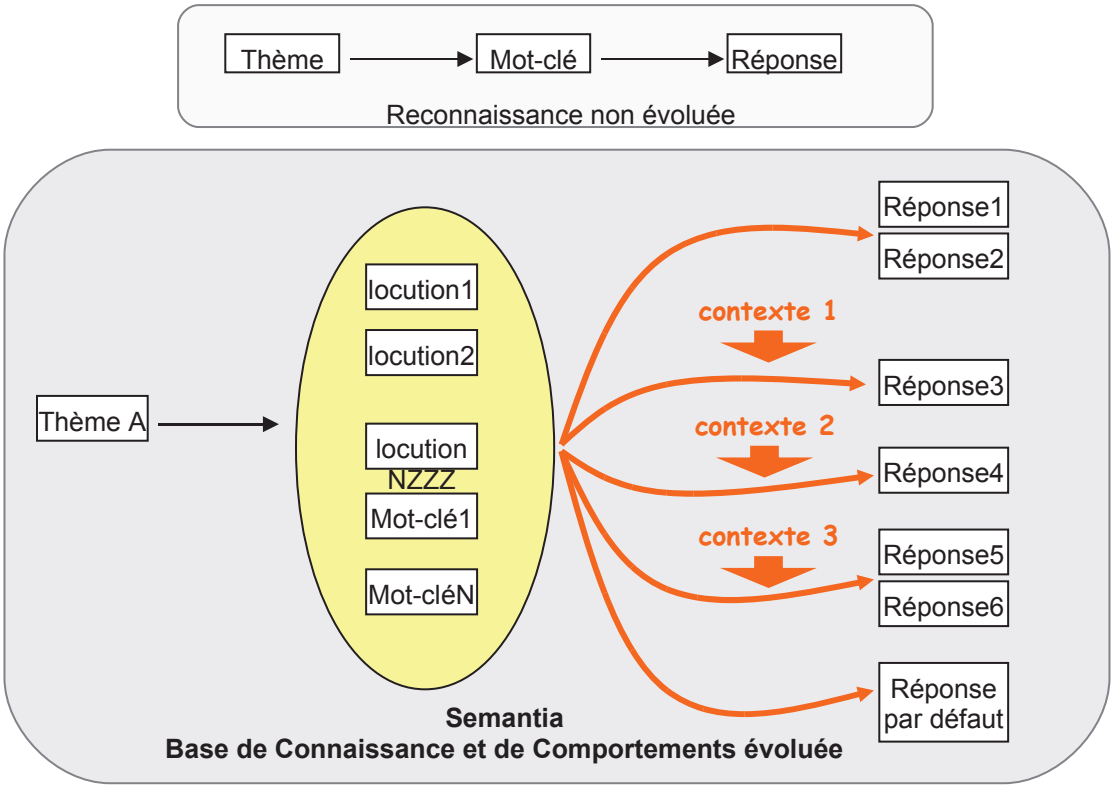

Fig. 3 : Fonctionnement de la détection de thèmes

Lorsqu'un utilisateur pose une question à un service Semantia, la base de connaissances qui le constitue reconnaît des mots-clés et des locutions, grâce à des **descripteurs** de type expressions régulières, et identifie ainsi des thèmes dans le message envoyé. Elle fournit une réponse en fonction du thème ou des thèmes qui viennent d'être extraits et du contexte de dialogue avec l'internaute (questions posées auparavant, identification dans un espace client...).

Dans le cadre de ce stage, il s'agit uniquement de l'extraction de thèmes car la base de connaissances sur laquelle nous nous appuierons et la base de connaissances que nous créerons ne seront pas utilisées pour une interface de type Web Agent. Les réponses renvoyées seront des couples attributs/valeurs (par exemple, *Hôtel Non Fumeur* = 'oui'). Les réponses ne seront donc ni plurielles, ni conditionnées par le contexte de dialogue et c'est la raison pour laquelle nous n'évoquerons pas en détail cette partie de la base de connaissances.

Au delà de la reconnaissance de locutions, la base de connaissances Semantia gère les synonymes et les différents langages (du plus formel jusqu'aux abréviations). Elle peut optimiser la détection de thèmes grâce à des associations entre ceux-ci. Par exemple, le thème *Réduction Réseau Continent Corse* résulte d'une association entre le thème *Les conditions de réduction* et le thème *Réseau Continent Corse*. Elle est également tolérante aux fautes de frappe et aux fautes d'orthographe.

2) Description détaillée : les lexiques

Cette vaste gestion des variations du langage est possible grâce à des lexiques. Ils sont inclus dans la base de connaissances, ce sont des dictionnaires d'équivalences. Ils se chargent de normaliser les messages envoyés ; cette standardisation permet d'éliminer les mots vides (déterminants, chiffres et nombres...), de simplifier des formes (singulier/pluriel, verbes conjugués...) ou de remplacer des mots ou groupes de mots par d'autres (mots longs/synonymes courts...). Nous nommons **générique** la forme qui remplace toutes les équivalences auxquelles elle correspond. Par exemple, toutes les **équivalences** *gîte*, *résidence*, *manoir* seront remplacées par le générique *hôtel* dans le texte reçu.

La base de connaissances peut contenir un seul ou plusieurs lexiques, le type de normalisation peut différer et il est alors plus clair de créer des lexiques distincts. Une base de connaissances peut, par exemple, être composée des lexiques suivants :

- Lexique par défaut ;
- Adverbes ;
- Lexique Métier.

B - La périphérie de la technologie : autour de la base de connaissances

Après avoir défini ce qu'était une base de connaissances, nous allons présenter l'organisation des services Semantia dans la technologie de l'entreprise.

1) Structurer l'information

Les services sont multi-sources et multi-canaux : le texte transmis à Semantia peut être une demande de renseignements, de diagnostic ou il peut nécessiter une modération, une structuration; sa provenance peut être multiple (fig. 4).

En passant dans la base de connaissances, l'information est nettoyée et réécrite. Cela permet une simplification ciblée de la demande afin de détecter plus facilement les thèmes présents. L'information est renvoyée sous forme d'information structurée : thématiques, couple attribut/valeur, réponse textuelle, identifiant …

Fig. 4 : Structurer l'information

2) Répondre par une action automatique

En fonction de la détection, la base de connaissances réagit par une réponse, éventuellement conditionnée par certaines variables conversationnelles ou relatives au profil de l'utilisateur, ou bien modérée par un cahier juridique...

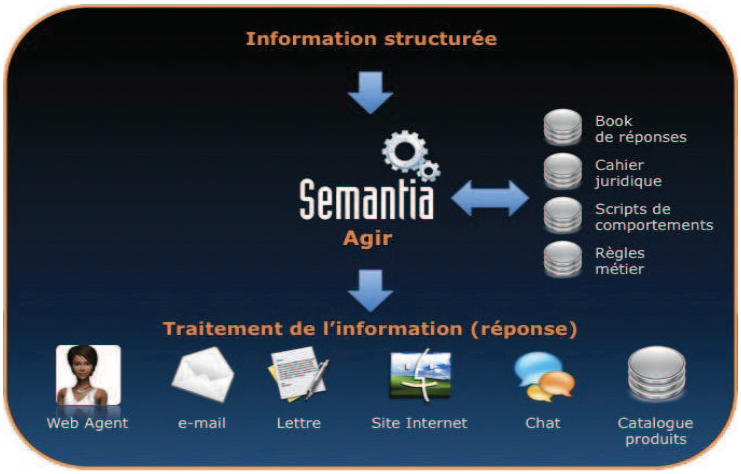

Fig. 5 : Répondre par une action automatique

C - Semantia : une approche plutôt applicative

Comme nous l'avons vu, Semantia travaille plus avec des règles spécifiques (ou règles métier) qu'avec des algorithmes génériques et déterministes.

Ses services disposent de bases de connaissances thématiques : le champ de compréhension est axé sur un domaine prédéfini. Certes, ces bases de connaissances ne prédisent rien et ne fonctionnent pas par apprentissage, mais leur objectif est purement applicatif et les services sont opérationnels et plus performants par ce procédé.

Pourtant, Semantia ne s'éloigne pas totalement de la recherche universitaire traditionnelle. Elle mène des projets de recherche et développement (R&D) régulièrement. Le dernier, en janvier 2010, en collaboration avec Pierre Gotab, a permis d'évaluer les bénéfices d'une éventuelle utilisation d'une technique d'*active-learning* et de classifieurs automatiques dans l'optique de réduire l'effort d'annotation et d'augmenter la qualité des bases de connaissances. Cette étude constitue un réel objet de recherche universitaire tout en étant orienté sur une application directe.

Nous avons détaillé l'histoire, le marché, les services et la technologie de Semantia. Dans ce contexte, l'objet d'études du projet a été explicité ; nous pouvons débuter la mise en place des deux solutions de traduction à tester.

Partie 2

Une base de connaissances par langue

Chapitre 1 – Pistes méthodologiques

La première partie du projet de stage est consacré à la réalisation d'une des solutions de traduction : la création d'une base de connaissances nouvelle pour une autre langue que le français.

Nous allons ici expliquer les réflexions menées et développer les pistes explorées pour mettre en place cette première solution. Nous détaillerons aussi la méthodologie choisie avant de présenter, dans le chapitre suivant, des exemples de traduction de la base de connaissances française (langue source) vers l'anglais (langue cible). Dans la dernière partie, nous conclurons sur les techniques employées pour la mise en place de cette première solution de traduction.

A - Comment traduire ?

1) Il n'y a pas qu'une traduction...

La linguistique théorique soulève beaucoup de questions sur la traduction, mais il n'en est pas moins reconnu qu'elle est possible en pratique à certaines conditions [Ruwet, N., 1964, p.141-144 et Woodsworth, J., 1988, p.2-3]. Il existe plusieurs façons de traduire un texte d'une langue à une autre, de la traduction littérale à la transposition. Ce sont les éléments de la base de connaissances, c'est-à-dire les intitulés des thèmes et les expressions qui permettent de les détecter - descripteurs et génériques – que nous allons transformer, et non des textes. Nous allons voir que même la traduction d'une simple expression de type descripteur peut soulever des questions.

2) Traduction littérale : le constat

Pour commencer, nous avions à disposition une base de connaissances en français sur l'hôtellerie. Notre première tâche a été de tester la traduction pure et simple vers l'anglais des descripteurs de celle-ci, à la façon d'une traduction littérale en littérature.

Cette traduction « test » n'a pas été transposée directement dans une nouvelle base de connaissances, nous pensions que cela nous ferait perdre du temps. Chaque traduction de descripteur a donc été répertoriée dans un tableau, contenant l'intitulé du thème général (ou catégorie) et du sous-thème (ou thème), le modèle du descripteur d'origine et la traduction correspondante. Ci-dessous se trouve un aperçu du tableau pour les thèmes *wheelchair-friendly* et *tour desk* :

CATEGORIE	THEME	DESCRIPTEUR FRANCAIS	TRAD DESCRIPTEUR ANGLAIS
Hotel services	wheelchair-friendly	ACCES HANDICAPE	HANDICAPED ACCESS
Facilities			
	tour desk	AGENCE DE VOYAGES	TRAVEL (INFORMATION) AGENCY
		BUREAU D'EXCURSIONS	TOUR([S|IST]) (INFORMATION) AGENCY
		VENTE DE BILLETS	EXCURSION(S) (INFORMATION) AGENCY

Fig. 6 : Tableau de correspondances descripteurs français/traduction anglaise

Nous nous sommes rapidement rendus compte que la traduction littérale offrait des possibilités limitées quant à la création d'une nouvelle base de connaissances.

La traduction littérale d'une expression française utilisée fréquemment dans l'usage pour un certain thème perdait de la pertinence pour une détection en anglais. La correspondance exacte des expressions (notamment des expressions figées) d'une langue à l'autre ne se vérifie pas souvent [Gawron-Zaborska, M., 2000, p.3 et Darbelnet, J., 2000, p.8]. Aussi, la décision d'un passage simple, par une traduction littérale, d'une base de connaissances monolingue à une autre ne nous est pas apparu satisfaisante.

3) Traduction à l'aide d'un corpus

En cherchant à trouver une meilleure traduction pour certaines expressions, nous nous sommes constitué un dossier de liens vers différents sites de réservations d'hôtels. Grâce à ces liens, nous pouvons considérer le contexte ou les locutions synonymes des expressions recherchées. Avec ce moyen de comparaison, nos descripteurs, au départ simplement traduits, ont été adaptés à l'anglais. Ils s'enrichissent d'un mot supplémentaire ou d'un synonyme plus utilisé – mot ou expression que nous retrouvions plusieurs fois -. Il a donc été décidé de reproduire cette démarche pour la traduction de chaque descripteur et pour rendre ce procédé plus performant, nous avons créé un corpus de fiches hôtel à partir du dossier de liens utilisé.

Nous avons réuni une cinquantaine de descriptions d'hôtels dans une base de données MySQL. Nous expliquons ci-dessous notre processus de sélection.

Aux premières sources utilisées, nous avons ajouté des fiches plus variées par la localisation et le classement des hôtels. Nous avons sélectionné des hôtels présents sur différents continents, mais nous nous sommes rapidement aperçu de l'uniformisation des services et des équipements. De plus, la plupart des sites d'hôtels non anglophones ont recours à des traducteurs automatiques pour obtenir un portail anglais, or ce n'est pas ce qui nous intéresse ici. Pour ce qui est du « standing », le corpus est composé d'une moitié d'hôtels de moins de quatre étoiles et d'une autre moitié d'hôtels de 4 étoiles et plus.

Ces fiches décrivant l'hôtel et ses services peut provenir directement de la page Web de l'hôtel, d'un site de référencement/réservations d'hôtels ou du site Internet d'une chaîne hôtelière : il n'était pas nécessaire de puiser les descriptions ailleurs que sur Internet, la structure et le contenu des fiches ne sont pas réellement différents sur des catalogues.

Nous devons obtenir assez de fiches hôtel pour disposer d'un panel large des types d'expressions ou de formules utilisées dans le domaine, ainsi que des types de services et équipements proposés aux clients.

Le tableau ci-dessous classe les fiches hôtel contenues dans notre corpus en fonction des sources et des lieux :

Localisation / Sources	Europe	Am. Nord	Am. Sud	Asie	Afrique	
Sites de réservation	17	7	2	2	-	**28**
Chaîne hôtelière	4	1	-	-	3	**8**
Hôtel isolé	10	2	-	-	-	**12**
	31	**10**	**2**	**2**	**3**	

Fig. 7 : Tableau de classement du corpus de fiches hôtel

Cette base de données hôtelières n'est certes pas exhaustive mais elle n'a pas pour but d'être un « corpus de référence » au sens où l'entend Sinclair, J., 1996, p.10. L'idée est d'avoir à disposition assez de matière pour travailler sans s'appuyer sur des données totalement biaisées, choisies uniquement pour leurs apports futurs à la traduction. En effet, il aurait été plus facile de ne s'appuyer que sur des descriptions de grands hôtels, - ceux-ci proposant plus de services que d'autres types d'hôtels, ils apportent plus de contenu pour former les descripteurs -. Mais nous avons opté pour une approche plus « scientifique » : « améliorer la représentativité du corpus […] [en précisant] la production et la réception de chacun de ses composants, en lien avec les motifs qui ont conduit à la création du corpus, mais aussi […] [en déterminant] sur des bases objectivables les différents emplois du langage auxquels on s'intéresse. » [Habert, B., 2000, p.5].

Notre base de données est donc plus apparentée à un échantillon des descriptions d'hôtels qu'à un corpus représentatif des fiches d'hôtels sur Internet. C'est sur la base de ce « corpus échantillon » que nous nous sommes appuyés afin d'adapter la traduction des descripteurs à la langue cible.

B - Méthodologie

Maintenant que nous savons avec quelles données travailler et comment traduire, nous pouvons décrire les étapes nécessaires à l'adaptation d'un thème de la base de connaissances d'une langue source à une langue cible.

Nous avons évoqué plus haut l'utilisation d'un tableau pour clarifier la correspondance entre les descripteurs français et les descripteurs anglais d'un thème. Tout en avançant dans notre réflexion sur la méthode de traduction, nous continuions à transcrire les descripteurs des thèmes. Par conséquent, une fois la méthodologie validée, la moitié des thèmes avaient déjà été traduits dans ce tableau, comme ci-dessous :

CATEGORIE	THEME	DESCRIPTEUR	EXTRAIT CORPUS
Hotel services	wheelchair-friendly	HANDICAPED [ACES|ACESES|ACESIBILITY|FACILITIES]	This hotel does not offer any accessibility features.
Facilities		REDUCED MOBILITY [ACES|ACESES|ACESIBILITY|FACILITIES]	Disabled Accessibility
		[DISABLE|DISABLED|DISABILITY|DISABILITIES] [ACES|ACESES|ACESIBILIT	Disabled Access
		WHELCHAIR([_|-])[ACES|ACESES|ACESIBILITY|ACESIBLE|FACILITIES]	Disable access(es)
		WHELCHAIR([_|-])FRIENDLY	Handicapped access
		[GUEST(S)|VISITOR(S)|PEOPLE] WITH [LIMITED MOBILITY|DISABILITIES]	Disabled Facilities Available
			Accessibility disabled persons
			Facilities for disabled
			disabled guests
			disabled travellers
			Wheelchair access
			Wheelchair-friendly
			Reduced mobility facilities
			guests with limited mobility
			Guests with Disabilities
			visitors with disabilities
			people with disabilities

Fig. 8 : Tableau de correspondances extraits du corpus/descripteurs traduits

Tous ces nouveaux descripteurs ont été reportés dans une base de connaissances créée pour l'anglais. Nous avons construit des lexiques car certains groupes d'équivalences étaient nécessaires à plusieurs thèmes. Par exemple, le générique *facility* et ses équivalences *facilities*, *access*, *accessibility, etc* ont dû être créés, ils sont employés par les thèmes *Wheelchair-friendly* (fig. 8) et *Business center.*

Pour la traduction de la deuxième moitié des thèmes, nous ne passons plus par le tableau. Cette étape n'est plus nécessaire étant donné que nous avons validé une méthodologie. Le passage du français à l'anglais pour une catégorie hôtellerie se fait donc directement dans la base de connaissances, en suivant les étapes ci-dessous :

- traduction simple des intitulés de la catégorie et des thèmes correspondants,
- évaluation de la possibilité d'une traduction littérale des descripteurs français,
- recherche dans le corpus des expressions se rapportant au thème,
- enrichissement des descripteurs traduits grâce aux extraits sélectionnés du corpus,
- si nécessaire, création ou amélioration de(s) lexiques concernés,
- tests de vérification de détection des catégories qui viennent d'être transcrites.

Cette succession de tâches n'est évidemment pas toujours suivie de manière exacte, les variations de langue et les défauts de l'échantillon de fiches hôtel nous ont parfois poussés à ajouter ou supprimer des phases dans cette méthodologie.

Chapitre 2 – Création d'une base de connaissances anglaise

Nous venons de voir les étapes nécessaires à la traduction de la base de connaissances, ces procédés sont soumis à la variation des thèmes et des expressions utilisées dans l'usage (ici, dans le corpus). Nous présentons dans ce chapitre quelques exemples concrets de traduction, du plus simple au plus complexe ; il s'agit d'un aperçu des questions soulevées par la création d'une base de connaissances traduite et par la modélisation des descripteurs et des lexiques.

A - Démonstration de l'application de la méthodologie : cas simple

Comme nous l'avons expliqué, la sélection des thèmes se fait par le repérage de locutions, normalisées grâce aux descripteurs. Ceux-ci prennent la forme d'expressions régulières s'il est nécessaire. Voici un exemple d'élaboration de descripteurs avec le thème *Renovated building* (catégorie *Hotel state*).

Nous commençons par une traduction simple du descripteur français, *hôtel (récemment) rénové* devient *hotel* (*recently) renovated.* Dans ce cas, il est bâti sur le modèle d'un nom suivi d'un adjectif, avec un adverbe facultatif placé entre eux. Ce qui est optionnel est entre parenthèses comme dans une expression régulière.

En consultant le corpus, nous observons que l'adverbe *recently* peut autant se trouver après l'adjectif qu'avant ; un nouveau descripteur est né : *hotel* (*recently) renovated* cohabite avec *hotel renovated* (*recently).*

Nous notons aussi que l'adjectif du descripteur possède quelques synonymes. Pour les prendre en compte, il suffit de créer un lexique : *renovated* en est le générique et *restored*, *refited*, *refurbished...* en sont les équivalences. Nous avons choisi l'option du lexique plutôt que celle de l'ajout direct dans le descripteur car le nombre d'équivalences est trop important, le descripteur risquerait d'alourdir le temps de traitement et serait moins lisible à l'écran. A l'inverse, nous ajoutons aux descripteurs les correspondances de l'adverbe *recently* car elles sont peu nombreuses et rien ne permet de supposer qu'un tel lexique sera réutilisable pour d'autres thèmes - contrairement au lexique *renovated* qui peut aussi concerner les chambres de l'hôtel ou des installations quelconques -. Suite à ces remarques, les descripteurs sont adaptés avec des crochets et des pipes qui permettent l'énumération d'éléments identiques : *hotel* ([*recently*|*tastefully*|*sensitively]) renovated* et *hotel renovated* ([*recently*|*tastefully*|*sensitively]).*

Néanmoins, il reste un modèle d'expression, tiré du corpus et apparaissant souvent, qui n'est pas encore inclus dans les descripteurs, il s'agit du modèle adverbe et adjectif seuls, comme dans *fully refurbished this winter*. La première étape est l'addition de *fully* aux correspondances de *recently* : *hotel* ([*recently*|*(taste)fully*|*sensitively]) renovated* - et il en sera de même pour le second descripteur - nous faisons l'économie d'un ajout de mot au descripteur en rendant *taste* optionnel. Ensuite, pour pouvoir détecter ce type de locution, l'élément *hotel* doit être facultatif dans le cas où il n'y a qu'un adverbe et un adjectif. Nous pouvons alors proposer les descripteurs suivants : *hotel renovated* et *[recently*|*(taste)fully*|*sensitively]) renovated.*

Pour finir, il faut tester les descripteurs créés et les réajuster s'ils présentent des anomalies. Nous n'avons pas rencontré de problèmes pour cet exemple.

B - Démonstration de l'application de la méthodologie : cas complexe

1) De *Description par adjectifs* à *Description de l'hôtel*

Dans la catégorie *Hotel description* (ou *Description de l'hôtel*), on peut trouver le thème *Romantic* dont les descripteurs ont été créés par traduction littérale et par expressions similaires. Une des expressions tirée du corpus, *privacy world*, nous donne à réfléchir sur la base de connaissances elle-même.

La catégorie contenant les thèmes *Romantic* ou *Comfortable* était nommée *Description par adjectifs* dans la base de connaissances française sur laquelle nous nous appuyons. Cependant, lors de l'adaptation, nous nous sommes aperçus de la perte d'informations qu'occasionnait un tel intitulé de catégorie puisque beaucoup d'éléments évoquant le confort ou le côté romantique d'un hôtel sont des noms ou des groupes nominaux. *Description par adjectifs* est devenu *Hotel description* et les descripteurs des thèmes comportent autant d'adjectifs que de noms. *Romantic*, *romantically*, *charm*, *charmed*, *privacy world* sont parmi les locutions choisies pour le thème *Romantic*. Si *privacy* est accompagné du nom *world*, c'est parce que, détecté seul, il pourrait avoir un tout autre sens, par exemple, s'il était suivi de *space* ; ce cas est d'ailleurs utilisé pour le thème *Non smoking rooms* lorsque *privacy space* est accompagné de *non smoking.*

La traduction du français vers l'anglais nécessite parfois l'adaptation de la base de connaissances même si l'on reste dans les mêmes catégories et les mêmes thèmes.

2) Enrichissement des thématiques

Nous venons de voir que la base de connaissances française avait parfois besoin de remaniements pour être plus performante, celle que nous créons en anglais est donc plus complète. Nous avons constaté que certaines locutions du corpus pouvaient servir à créer de nouveaux thèmes.

Certains hôtels disposent de jacuzzis, de saunas et de hammam, nous souhaitons donc ajouter les descripteurs correspondants - *jacuzzi(s)*, *sauna(s)* et *hammam(s)* - dans la base de connaissances anglaise. Faut-il les ajouter dans le thème *Swimming pool* ? Nous avons préféré créer un nouveau thème *Beauty services* dans la catégorie *Hotel facilities*, d'autres descripteurs tels que *manucure* ou *massage(s)* sont allés étoffer ce thème.

De la même manière, *Wedding service* a été créé : nous rencontrons des difficultés à trouver d'autres descripteurs que *bridal suite* pour le thème du même nom (catégorie *Hotel rooms*), car les locutions qui se rapportent au mariage étaient trop générales pour un thème aussi réduit. *Wedding service* nous permet de détecter plus d'éléments que *bridal suite* tout en gardant la même idée thématique.

C - Démonstration de l'application de la méthodologie : tests

La dernière étape de la méthodologie donnée plus haut est la phase de tests des descripteurs. Nous avons choisi d'effectuer les tests au fur et à mesure de la traduction des thèmes. Pendant les tests, l'inadéquation de certains descripteurs a été révélé et leur construction a dû être revue.

1) Polysémie : le problème des mots-clés

En vérifiant la catégorie *Sports et leisure*, nous nous sommes rendus compte que le thème *Swimming pool* était parfois détecté alors que l'hôtel ne disposait pas de piscine. Des descripteurs comme le mot clé *pool* induisait en erreur la base de connaissances et nous allons expliquer pourquoi.

D'après nos observations sur le corpus, il est d'usage d'employer les mots *pool* ou *swimming* seuls pour désigner une piscine ou le fait que les clients auront la possibilité de nager. Le problème est qu'aucun de ces mots ne peut être un mot-clé ; en effet, *swimming* peut être mentionné pour évoquer la mer proche et *pool* est un mot polysémique en anglais, par métonymie : il peut désigner une piscine (*swimming pool*) ou un billard (*pool table*). Afin de disposer de plus de descripteurs que le simple *swimming pool*, nous avons créé des descripteurs trop larges. Plusieurs solutions sont envisageables pour améliorer la détection. Nous pourrions purement et simplement supprimer le mot-clé qui induit le moteur en erreur, mais avant d'arriver à cette solution, nous allons regarder dans le corpus si il ne peut pas être enrichi pour devenir une locution efficace.

En considérant les contextes des occurrences du mot *pool* dans le corpus, nous avons trouvé comment remédier à ces fautes de détection. Grâce à des expressions comme *non heated outdoor pool* ou *indoor pool,* le mot-clé a été complété, le descripteur ainsi créé ne détecte plus « trop large » ; et les expressions *pool/snooker* et *pool table* pour le sens de billard sont alors susceptibles d'être détectées par ailleurs.

Nous retrouvons le même phénomène avec le descripteur *walk(ing)* dans le thème *Sports activities* (toujours dans la catégorie *Sports et leisure*). La détection du thème est enclenchée par des expressions du type *walking distance…* ou bien *Located just a few minutes' walk from Piccadilly [...],* or de telles phrases n'évoquent pas la possibilité de randonnées ou de balades autour de l'hôtel. Contrairement au cas précédent, nous n'avons pas trouvé de locutions dans le corpus susceptibles d'enrichir le mot-clé ou d'éliminer l'ambiguïté.

Les expressions utilisant *walk* ou *walking* dans un autre sens que randonnée sont trop nombreuses pour être normalisées dans un descripteur performant. De plus, il faut savoir que *Sports activities* est un thème général qui regroupe toutes les expressions les plus rencontrées sur les activités sportives dans les fiches hôtel, qui n'ont pas de thèmes spécifiques dans la catégorie *Sports et leisure*. La suppression d'un mot-clé parmi tous les descripteurs ne nuit donc pas vraiment à la détection du thème, au contraire, elle permet d'éviter des erreurs. En conséquence, nous avons simplement supprimé *walk* des descripteurs du thème *Sports activities*.

Suite aux tests, les mots-clés et les descripteurs de la catégorie *Sports et leisure* sont donc modifiés afin de rendre la détection des thèmes de celle-ci plus fine et plus pertinente.

2) Enrichissement d'un mot-clé non polysémique

Lorsque des thèmes évoquent des concepts et non des services matériels, l'élaboration des mots-clés et des descripteurs est plus complexe.

Prenons l'exemple du test sur un thème comme *refined* (raffiné) dans *Hotel description* ; parmi ses locutions se trouve le mot-clé *tasteful*, ce qui paraît tout à fait pertinent. Pourtant, en contexte, cet adjectif peut aussi faire référence à la qualité de la nourriture du restaurant de l'hôtel ou il peut avoir un rapport avec une information touristique, un met culinaire traditionnel du lieu où se situe l'hôtel... Or, notre thème évoque plutôt le raffinement de la décoration, du mobilier et de l'aspect général de l'hôtel. *Tasteful* paraît être un mot-clé idéal puisque que ses emplois sont multiples et qu'il garde malgré tout le sens de quelque chose de distingué; mais ces significations ne correspondent pas au thème.

Tout comme dans l'exemple précédent, nous avons examiné le corpus avant de prendre la décision de supprimer le mot-clé. Il était possible de l'étoffer à l'aide de noms tels que *style* ou bien *decor*. Mais certains éléments qu'il paraît possible d'ajouter peuvent poser problème, par exemple, nous avons choisi d'écarter *world* pour éviter de détecter des expressions comme *in our restaurant, we will discover a tasteful world.* Le mot-clé *tasteful* est transformé en descripteur plus performant.

3) Nettoyage du message source : le contre-lexique

Dans les exemples précédents, nous avons corrigé des erreurs de reconnaissance dues à des descripteurs pas assez précis ou ambigus par l'ajout de précisions. Mais il existe des cas où cette méthode n'est pas envisageable. Nous allons voir que parfois la base de connaissances doit être modifiée pour agir directement sur le message transmis et non pour affiner les éléments qui la composent.

Dans la phrase *[...] you want to sample the infinite charm of the Luxembourg gardens,* l'élément *Luxembourg gardens* fait remonter le thème *With garden/patio/terrace* de la catégorie *Hotel kind*, alors qu'il s'agit de la proximité d'un jardin public et pas la présence d'un hôtel avec jardin. Là encore, les solutions disponibles sont l'enrichissement du mot-clé ou sa suppression. Mais dans ce cas, même avec l'aide du corpus, nous ne découvrons aucun moyen d'enrichir le mot-clé *garden(s)* et il n'est pas envisageable de le supprimer puisqu'il est le seul à déclencher la détection du thème *With garden/patio/terrace.* Nous optons donc pour une troisième issue : la création d'un lexique supplémentaire.

Ce lexique appelé *contre-lexique* va nous permettre de supprimer *Luxembourg gardens* du message à thématiser, il est placé en tête des lexiques afin de s'assurer qu'il sera le premier à œuvrer. De cette manière,

La solution du *contre-lexique* a été pérennisé par sa réutilisation dans de nombreux thèmes. Par exemple, *telephone line* de la catégorie *Hotel rooms* : dès qu'une fiche hôtel renseignant le numéro du standard téléphonique de l'établissement, la base de connaissances renvoyait le thème (comme dans cet extrait du corpus : *by phone at 512-474-5911*). Nous avons donc placé dans le *contre-lexique* un descripteur du type mot-clé téléphone suivi de nombres.

A l'aide de ces exemples, nous voyons les types de changements opérés sur la base de connaissances suite aux tests dans le but d'obtenir une meilleure détection.

Chapitre 3 – Premiers résultats

La mise en place de la première solution nous permet de proposer quelques réflexions sur la méthode utilisée et sur la suite du projet d'évaluation.

A - De l'utilité du corpus

Le corpus a été utile à l'adaptation de la base de connaissances française à l'anglais, mais il a aussi été la source d'erreurs.

1) Des avantages importants…

Le thème *Family rooms*, de la catégorie *Hotel rooms*, va nous permettre de visualiser un exemple typique de la façon dont a été adaptée la base de connaissances.

Pour ce thème, la base de connaissances française ne proposait aucun lexique et qu'un seul descripteur : *chambres familiales.* Nous ne trouvons que peu de fois l'expression *family rooms* dans le corpus de descriptions d'hôtels. Pour l'adaptation en anglais, il a donc fallu déceler quelles autres expressions montraient que la présence d'enfants était possible, sans avoir explicitement *family rooms*. Or, dans certaines fiches hôtel, des éléments comme *cots on request*, *crib in room* ou *infant's bed available* sont présents, nous les avons donc ajoutés comme descripteurs.

L'avantage de s'appuyer sur un corpus de descriptions d'hôtels et donc sur des éléments en « contexte » est mis en lumière, les descripteurs deviennent plus pertinents pour la langue cible.

Des adaptations encore plus poussées peuvent être faites pour repérer la possibilité d'obtenir une chambre familiale dans un hôtel grâce à l'ajout de lexiques : *cots on request* et *crib in room* sont des expressions formées de manière identique, un nom suivi d'un groupe adjectival. Les trois éléments *cots*, *crib* et *infant's bed* sont équivalents et peuvent donc être regroupés dans un même lexique. *On request*, *in room* et *available* sont aussi des expressions similaires, elles sont susceptibles de constituer un même lexique.

Nous obtenons donc un nouveau lexique que nous appellerons *crib* contenant *cot* et *infant's bed* ainsi que leurs correspondants au pluriel. La création d'un lexique avec les groupes adjectivaux et *available* ne nous a pas paru pertinente, celui-ci n'aurait pas été réutilisé ni dans d'autres thèmes, ni par d'autres lexiques, et il n'était pas nécessaire à la lisibilité des descripteurs.

L'adaptation de la base de connaissances passe par une traduction littérale et/ou en contexte des descripteurs d'un thème du français et surtout par leur enrichissement, mais elle entraîne aussi la création de nouveaux lexiques.

2) … mais aussi des inconvénients.

Nous venons de voir avec le thème *Family rooms* l'importance du corpus dans la comparaison des traductions littérales et des usages. Il y a des cas où les expressions à traduire (ou des expressions similaires de même sens, comme pour l'exemple précédent) sont peu présentes en contexte. Il apparaît alors nécessaire d'utiliser un dictionnaire généraliste pour adapter le thème à l'anglais et avoir assez de descripteurs pour le détecter. Ce défaut de données peut être autant dû à la composition de notre corpus qu'à la faiblesse de l'utilisation du thème en question dans une description d'hôtel.

Nous avons été confronté à ce problème avec le thème *Old building* de la catégorie *Hotel state*, la base de connaissances française faisait appel à des descripteurs de type *hôtel datant du 15ème siècle*, *hôtel du début du 16ème* or ces constructions sont effectivement peu représentées dans les fiches hôtel anglaises. En plus des traductions littérales (*hotel dating from 15th century*), nous avons enrichi le thème de constructions typiquement anglaises comme *early 16th century hotel*. Ces syntaxes ont été trouvées sur des sites d'agences immobilières décrivant des biens anciens en vente. Peu de descriptions d'hôtels (même hors corpus) contiennent des expressions sur la date des bâtiments. Évidemment, ces descripteurs ont dû être dérivés pour tous les siècles concernés.

L'adaptation de ce thème à l'anglais nous montre les limites de l'utilisation d'un corpus, mais celui-ci nous a tout de même permis de créer de nouveaux lexiques avec des équivalences de *old*, par exemple, comme *ancient*, *historic*, *vintage* ou bien *edwardian*.

A la lumière du travail déjà effectué, la linguistique de corpus semble une solution intéressante. La traduction a été facilitée grâce à cette méthode. De plus, ce fonctionnement a aussi mis en valeur la possibilité de regrouper certains thèmes et la nécessité d'en créer de nouveaux. Les lexiques pour l'hôtellerie vont donc être remaniés et cela sera, sans doute, le cas à chaque adaptation de la base de connaissances à une nouvelle langue.

La possibilité d'avoir à notre disposition un corpus plus exhaustif ou plus important n'aurait, à notre avis, pas amélioré de façon significative les résultats : les thèmes peu représentés dans le corpus le sont visiblement sur tout le Web. En effet, parfois nous avions des difficultés à enrichir un mot-clé trop « gourmand » ou à sélectionner des synonymes pour élargir la détection, et dans ces cas-là, une recherche sur un moteur était aussi fastidieuse qu'une fouille du corpus (cf. exemple de la construction des descripteurs de dates de bâtiments, plus haut).

Les défauts de celui-ci nous paraissent donc mineurs par rapport aux avantages apportés par son utilisation.

B - Retours des tests

Dans le chapitre précédent, nous présentions des problèmes de détection dus à des mots-clés.

Il faut savoir que les tests ne révèlent pas seulement la nécessité de changement dans la base de connaissances, mais aussi des erreurs dans les descripteurs. Nous n'avons pas souhaité donner plus de détails puisqu'il s'agissait le plus souvent d'oublis de prise en compte du pluriel d'un mot ou de signification de la possibilité de présence d'un déterminant ou d'un adverbe.

Le détail de ces modifications n'apporte pas d'éclairages pertinents sur la mise en œuvre de cette première solution de traduction.

C - Évaluation : étape du tableau de correspondances

Nous l'avons expliqué dans la partie Méthodologie du chapitre 4, la première moitié de la base de connaissances française a été traduite avec un tableau intermédiaire. Cette étape supplémentaire dans le déroulement de la méthode de traduction doit être prise en compte si l'on veut que les mises en place des deux solutions soient comparables. Il faudra soit utiliser le même procédé, soit comptabiliser cette différence (et la durée de son application) pour l'évaluation.

Pour nous permettre d'évaluer cette traduction humaine à une traduction automatique, il nous faut mettre en place la deuxième solution.

Partie 3

Base de connaissances multilingue

Chapitre 1 – Pistes méthodologiques

La seconde solution de traduction à tester dans ce projet consiste à ne disposer que d'une seule base de connaissances en français et à exploiter un système de traduction automatique pour rendre les services Semantia multilingues. Pour pallier les erreurs d'interprétation, nous allons créer un lexique de règles dans la base de connaissances, sa construction est l'objet de la deuxième partie du stage. Dans ce chapitre, nous allons tout d'abord décrire comment nous avons rassemblé les données dont nous avions besoin; puis nous détaillerons la méthodologie employée pour mettre en œuvre le lexique. Nous verrons ensuite des exemples de création du lexique, dans le chapitre suivant. Enfin, nous ferons quelques remarques générales sur la mise en œuvre et l'évaluation de la solution de traduction.

A - Quels sont les données nécessaires ?

1) Une base de connaissances d'appui

L'équipe linguistique de Semantia souhaite disposer d'une base de connaissances française capable de traiter des messages traduits par un système automatique du commerce. La base de connaissances doit donc contenir un lexique prévu pour corriger les éventuelles erreurs du système de traduction automatique. Nous allons le créer mais il faut savoir dans quelle base de connaissances il va être ajouté.

Notre idée était celle d'utiliser comme appui la base de connaissances française sur l'hôtellerie que nous avons traduit en anglais dans la première solution. Mais la base de connaissances anglaise ainsi créée est plus riche que celle qui nous a servi de base en français. Nous avons vu qu'il était régulièrement nécessaire de changer un thème ou d'en créer un, avec tous les renforcements de descripteurs que cela implique (nous ne prenons pas en compte les modifications faites sur les lexiques, ceux-ci étant réellement dus à la transcription d'une langue à une autre). Les thèmes nouvellement créés et les modèles de descripteurs correspondants pourraient s'avérer utiles dans toutes les langues.

Ainsi, en accord avec les équipes de Semantia, la décision a été prise de « mettre à niveau » la base de connaissances française en s'appuyant sur les changements effectués en anglais : intégration des thèmes récemment ajoutés et adaptation de leurs descripteurs, nouvellement créés ou simplement enrichis. Nous employons le terme *adaptation* pour les descripteurs car ils n'ont évidemment pas été retranscrits tels quels, étant donné qu'ils sont en anglais ; mais leur structure, leur syntaxe, leur construction ont été réutilisés et réadaptés pour le français. Ils ont servi de modèle à l'enrichissement de la base de connaissances française.

Cette « mise à jour » a duré une quinzaine de jours. A partir de la nouvelle base de connaissances, nous pouvons enclencher la seconde solution de traduction. Le lexique pour les systèmes de traduction automatique va être créé à l'intérieur de celle-ci.

2) Un lexique de règles correctrices

Les règles qui corrigeront les erreurs les plus fréquentes du système de traduction automatique sont regroupés dans un lexique de la base de connaissances. Celui-ci permettra la transformation d'une erreur de traduction en mots-clés ou en descripteur permettant la détection du thème correspondant.

Nous avons vu plus haut (Partie 1, Chapitre 3) que les lexiques pouvaient être hiérarchisés, cette fonction va être utilisée pour cette solution de traduction. Le message envoyé passera d'abord par le lexique que nous allons créer. Les expressions sources d'erreurs seront donc les premières à être transformées après une traduction automatique. Ensuite, la base de connaissances fera intervenir les notions de génériques et équivalences habituelles dans l'extraction des thèmes.

Pour plus de lisibilité dans les explications qui vont suivre, nous nommerons le lexique relevant les erreurs principales de traduction, le « lexique traduction automatique » (ou « lexique TA »), et les lexiques déjà présents dans la base de connaissances seront appelés « lexiques d'appui ».

Pour réaliser cette deuxième méthode, nous disposons donc de la base de connaissances française sur l'hôtellerie, qui servira de pilier, et nous devrons créer en plus le lexique TA. Les questions qui se posent maintenant sont celles de la manière de relever les plus fréquentes erreurs que le système de traduction automatisé engendre dans le domaine de l'hôtellerie ainsi que celles du système de traduction que nous allons utiliser.

3) Un corpus pour relever les erreurs de traduction automatique

Nous allons d'abord étudier de quelle façon nous pouvons relever les erreurs les plus fréquentes d'un système de traduction automatisé, nous verrons ensuite quel système choisir.

Les étapes suivies pour la création du lexique TA sont différentes de celles de la première méthode. Il ne s'agit plus de traduire ou d'adapter les descripteurs des thèmes à une langue cible, mais de repérer les éventuelles erreurs de traduction de fiches hôtel pour un thème donné. Cependant, les différences de méthodologie sont pondérées par la nécessité d'agir d'une manière équivalente : les deux solutions doivent rester comparables pour être évaluables.

L'utilisation d'un corpus de fiches hôtel traduites automatiquement paraît nécessaire autant pour la comparaison des méthodes que pour la création du lexique TA. Nous disposons déjà, grâce à la création de la première solution de traduction, d'un corpus de fiches hôtel. Ce corpus est en anglais et nous avons besoin de fiches hôtel traduites automatiquement vers le français.

Il nous suffisait donc de traduire entièrement le corpus anglais vers le français à l'aide du service en ligne de Systran. Chaque fiche hôtel traduite a été insérée dans la base de données MySQL accueillant déjà le corpus en anglais, - une table prévue à cet effet a été ajoutée -. De plus, nous avons mis en place une interface PhP permettant une lecture aisée de la base de données, grâce à la visualisation côte à côte de deux versions d'une fiche hôtel (la fiche anglaise et sa traduction en français depuis un système automatique), . Le repérage des erreurs de traduction pour un ajout dans le lexique traduction automatique (ou lexique TA) est ainsi facilité.

4) Un système de traduction performant

Pour cette solution de traduction, nous disposons maintenant de toutes les données dont nous avons besoin. La dernière étape concerne donc le choix d'un traducteur automatique.

Premièrement, il faut nous interroger sur la nature des systèmes de traduction automatique. Il s'agit d'applications informatiques grâce auxquelles un utilisateur obtient automatiquement et rapidement (souvent, instantanément) une traduction de textes d'une langue source à une langue cible. Le texte peut être de différents types : lettres, rapports, articles, sites web, courriers électroniques, messages instantanés... Les systèmes de traduction automatique peuvent prendre la forme de logiciels (œuvrant de manière indépendante ou bien intégrant des fonctionnalités de traduction dans une autre application), mais aussi de services en ligne librement accessibles ou payants.

Pour avoir une meilleure compréhension de leurs capacités, nous pouvons aussi expliquer rapidement leur fonctionnement. Les systèmes de traduction automatique utilisent principalement deux méthodes : linguistique et/ou statistique. La méthode linguistique applique une analyse morphosyntaxique sur les textes à traduire et s'appuie sur des règles linguistiques pour retranscrire les éléments dans une autre langue, tandis que la méthode statistique se base sur des modèles probabilistes de traduction et des bases de données de la langue cible.
Quelques traducteurs automatiques emploient ces méthodes conjointement, nous parlons alors de méthode ou d'approche hybride.

Les services Semantia se doivent de traiter un grand volume de données de la manière pertinente et la plus rapide possible. Le système de traduction automatique que nous allons choisir doit prendre en compte ces besoins, car de lui dépend la performance des services proposées par l'entreprise.

Nous avons cherché parmi les systèmes de traduction automatique les plus réputés et trois on été retenus : le service Google Traduction, la solution de Softissimo (appelée Reverso) et la solution de Systran.

Google Traduction ne proposant pas ses services de traduction aux entreprises, nous nous sommes tournés vers les deux restants. L'objet du projet n'est pas la comparaison de la fiabilité et de la pertinence de solutions commerciales de traduction automatique, nous avons donc choisi en fonction d'arguments extérieurs.

La solution Systran propose un peu plus de possibilités de traduction (cinquante-deux paires de langues contre une vingtaine pour Reverso). Elle est performante lorsque les données à traduire sont d'un même domaine; or, ce sera quasiment toujours le cas pour une base de connaissances Semantia (cf. partie 1, chapitre 3 Semantia : une approche plutôt applicative).

Nous utilisons alors le service de traduction automatique en ligne de Systran pour traduire le corpus anglais en français. Après cette manipulation, nous disposons de toutes les données et de tous les outils pour commencer la conception du lexique TA.

B - Méthodologie

Cette deuxième solution de traduction nous permet d'aborder les bases de connaissances Semantia d'un autre point de vue. La manière de procéder diffère puisque l'adaptation dans le moteur se fait par le lexique et non par les locutions et les descripteurs. Les phases de réalisation de la méthode du lexique TA sont totalement dépendantes de ce qui se trouve dans le corpus, la base de connaissances reprend les formes des erreurs de traduction.

Si les étapes sont différentes, la conception des règles correctives est proche de l'adaptation des descripteurs de la première solution de traduction testée : pour un thème donné, nous élaborons des expressions régulières (mots-clés, descripteurs ou équivalences) en fonction d'éléments repérés dans un corpus. Dans le cadre de la seconde méthode, seules les équivalences nous intéresseront et pour répertorier les erreurs de traduction automatique les plus fréquentes dans un lexique, nous allons donc adopter la même façon de travailler que pour la première solution de traduction.

Pour une moitié du corpus, une phase de repérage grâce à un classement des erreurs dans un tableau est mise en place. Ce tableau fait écho à celui créé pour la traduction d'une base de connaissances du français à l'anglais - dans le but de faciliter la comparaison, nous avons utilisé à nouveau un tableau de correspondances pour la première moitié du corpus, nous reviendrons sur cette étape pour l'évaluation (Partie 4) -.

	CORPUS	SYSTRAN
FH N°1		
	Money exchange 24x24 hrs	Échange d'argent
	Wheelchair access	Accès de fauteuil roulant
	Pets (dog and cat) welcome	Bienvenue d'animaux familiers (chien et chat)
	Soft drinks bar	Barre de boissons non alcoolisées

Fig.9 : Tableau de correspondance extraits du corpus/extraits du corpus traduit

Les erreurs de traduction de la deuxième partie du corpus sont enregistrées directement dans la base de connaissances. Elles vont être normalisées (sous forme d'équivalences) pour intégrer le lexique TA. Les équivalences créées vont donc être transformées (en leur générique) et vont pouvoir renvoyer des thèmes. Il faut noter qu'au moment de la création d'une équivalence, les descripteurs du thème concerné doivent être vérifiés. Il se peut qu'ils fassent appel à des lexiques qui se répercutent sur l'équivalence, ou bien il est parfois possible d'avoir une simple modification à effectuer sur le descripteur pour que l'erreur de traduction soit acceptée et détectée. Nous verrons dans le chapitre suivant que cela peut poser problème.

Pour résumer, les étapes à suivre pour la mise en œuvre du lexique TA sont les suivantes :

- repérage des erreurs de traduction pour un thème donné,
- examen des descripteurs et des lexiques attenants,
- modification directe d'un descripteur ou d'un générique du thème, si nécessaire
- ajout d'une équivalence pour le thème dans le lexique TA, si nécessaire.

Dans le chapitre suivant, quelques exemples, nous montrerons comment cette méthodologie est appliquée sur le corpus.

Chapitre 2 – Enrichissement multilingue d'une base de connaissances : ajout de l'anglais

Nous venons de voir les étapes nécessaires à la conception d'un lexique TA. Dans ce chapitre, nous montrons comment ces phases se déroulent pour la mise en place la solution de traduction.

A - Démonstration de l'application de la méthodologie

Reprenons chacune des phases de la méthodologie dans un exemple concret afin de visualiser le parcours nécessaire à la future détection d'une expression contenant une erreur de traduction automatique.

Nous allons prendre l'exemple du premier thème de la base de connaissances *Hôtel non fumeur*, dans la catégorie *Services de l'hôtel.* Tout d'abord, nous relevons, pour ce thème, les différentes expressions contenant une erreur de traduction dans le corpus : *salles non fumeuses*, *salles non de tabagisme*, *non fumeur seulement*, *établissement sans fumée* et *propriété sans fumée*.

La seconde étape consiste à observer les descripteurs et des lexiques attenants au thème. Nous devons vérifier la possibilité d'inclure des modifications dans les descripteurs du thème directement et observer le contexte dans lequel nous allons opérer les changements pour ne pas entraîner des erreurs dans la détection. Pour le cas d' *établissement sans fumée,* nous remarquons, dans le lexique d'appui, l'existence de *sans* comme équivalence de *non*, et d'*etablissement* comme équivalence d'*hotel ;* - rappelons que tout comme les descripteurs, les génériques et les équivalences sont comparables à des expressions régulières et ne comportent donc pas d'accentuation -.

La méthodologie nous propose donc, en troisième lieu, de procéder à une modification directe sur un descripteur ou un générique du thème. A partir des expressions contenant des erreurs relevées dans le corpus, notamment *propriété sans fumée*, nous notons que l'élément *propriété* pourrait être intéressant à ajouter comme équivalence supplémentaire d'*hotel*. Il s'agit de la seule opération directe que nous pouvons effectuer dans la base de connaissances d'appui.

Enfin, il nous reste l'option d'un ajout d'équivalence dans le lexique TA. Nous disposons d'assez d'expressions pouvant être normalisées qui sont des erreurs de traduction automatique, nous créons donc un générique *hotel_non_fumeur* ayant pour équivalences *salle(s)_non_fumeuse(s)*, *salle(s)_non_a_tabagisme*, *hotel_non_fumee* et *propriete_non_fumee*. Comme nous l'avons vu, les équivalences *etablissement* et *sans* ont été remplacées par les génériques *hotel* et *non.*

Nous avons suivi, point par point, les étapes de la méthodologie donnée plus haut pour le thème *Hôtel non fumeur.* La détection de ce thème a été enrichie du générique *hotel_non_fumeur* dans le lexique TA. L'utilisation d'un système de traduction automatique avant le passage dans la base de connaissances apparaît comme une méthode réalisable.

B - Difficultés soulevées par l'imbrication des lexiques

Lorsque de l'ajout d'éléments aux lexiques déjà créés, il faut prendre garde à ne pas établir une équivalence qui existe déjà, cela créerait une boucle lors de l'appel de la base de connaissances sur l'élément transformé plusieurs fois. Or, les questions d'imbrication des lexiques sont parfois difficiles à appréhender. Si l'on crée des génériques de noms généraux, tels que *centre*, *chambre* ou *hotel*, les répercussions sont importantes sur tous les descripteurs de la base de connaissances. Un oubli de remplacement du nouveau générique ou de prise en compte d'une équivalence peut fortement détériorer la détection d'un thème.

Par exemple, nous avons été confronté à l'imbrication de plusieurs lexiques les uns dans les autres lors de l'adaptation du thème *Chambres familiales* au lexique TA.

Pour le thème *Salle de sports* de la catégorie *Sports et leisure,* le générique *centre* a été créé dans le lexique d'appui, il fait appel à des équivalences comme *centres*, s*alle(s)*, *club(s)* ou *espace(s)*. Des locutions comme *salle de sport, clubs de sport* remontent donc le thème concerné grâce à un seul descripteur, *centre_de_sport*. Par ailleurs, le générique *centre* est utilisé par d'autres thèmes. Par exemple, le thème *Salle de conférence* de la catégorie *Services de l'hôtel* avec le descripteur *centre_de_conference* qui permet de détecter des locutions comme *espace de conférence* ou *centre de conférence.*

Mais l'adaptation de la base de connaissances aux erreurs de traduction automatique va compliquer la tâche de ce générique.

Nous repérons dans le corpus des expressions pour le thème *Chambres familiales* : *salle de famille*, *chambre de famille* ou *pièce de famille*. Un ajout direct dans les descripteurs n'est pas envisageable, la question se pose donc sur les génériques et les équivalences à créer. Notre première réflexion nous a poussé à créer un générique *chambre* dans le lexique TA ayant pour équivalences *salle* et *pièce.* Mais la conception de ce lexique crée un souci dans la reformulation du texte, puisque *salle* est tour à tour transformé en *centre* (lexique d'appui) ou en *chambre* (lexique TA).

Nous cherchons donc à spécifier notre générique en le nommant *chambre_familiale,* il correspondant à *chambre_de_famille*, *piece_de_famille* et *centre_de_famille* directement transcrit de *salle de famille.* En poussant notre réflexion plus loin, nous réalisons que le lexique TA peut être rentabilisé en ajoutant un générique. Nous nous appuyons sur le lexique habituel de la base de connaissances en gardant l'idée d'un générique *centre* mais nous lui donnons comme équivalence le terme *pièce*. Ainsi, nous obtenons deux nouveaux génériques dans le lexique TA, un spécialisé *chambre_familiale* (équivalences : *centre_de_famille* et *chambre_de_famille*) et un généraliste *centre* (équivalence : *piece*). Le lexique d'appui contenant déjà la correspondance *centre*/*salle*.

L'ajout de génériques dans le lexique prévu pour la traduction automatique est dépendant des autres éléments de la base de connaissances française. Cela rend parfois complexe la mise en place de cette seconde solution.

C - Stratégie des éléments minimaux

Dans certains cas, les expressions contenant des erreurs de traduction automatique contiennent beaucoup de termes, contrairement à celles que nous venons d'étudier. Pour ne pas trop alourdir la base de connaissances, des stratégies sont mises en place.

Pour le thème *animaux acceptés*, toujours dans la catégorie *services de l'hôtel*, nous découvrons deux expressions intéressantes :
1- [...] souhaite la bienvenue a vous et a vos meilleurs amis quadrupèdes;
2- vous accueille et vous a quatre pattes meilleurs amis.

Pour des expressions aussi longues, l'idée est de chercher à garder le minimum d'éléments pertinents. La détermination de la réduction minimale dépend de l'existence d'un thème antonyme ou proche. Ici, la présence du thème *animaux non acceptés* dans la base de connaissances d'appui ne permet pas de réduire les expressions à *meilleurs amis quadrupèdes* et à *quatre pattes meilleurs amis* sans créer une ambiguïté. Par conséquent, dans la première, les deux premiers mots (*souhaite la*) peuvent être supprimés, tandis que dans la seconde, il n'y a pas de parties réductibles.

Pour continuer dans le sens de la méthodologie, nous cherchons de possibles modifications directes sur un descripteur ou un générique du thème. L'expression que nous avons raccourcie peut pallier des erreurs de traduction et est valable pour la base de connaissances française. Elle est donc ajoutée au thème concerné en tant que descripteur.

A l'inverse, la seconde expression est difficilement intégrable dans la base de connaissances française, elle va donc compléter les équivalences du générique *animaux acceptés* dans le lexique pour la traduction automatique, avec la forme suivante : *vous_accueille_et_vous_a_quatre_pattes_meilleurs_amis.*

Il est intéressant d'enrichir la base de connaissances française, quitte à adapter des erreurs afin de les transformer en descripteur. En effet, il nous semble plus performant d'avoir à disposition une base de connaissances solide en français, puisqu'elle est notre unique appui pour le lexique TA.

D - Erreurs de traduction sans conséquences pour la détection

La stratégie de réduction des expressions que nous venons d'évoquer a mis en lumière la présence de cas où une erreur de traduction faite par le système ne pose pas de problème dans la détection du thème. Nous avons constaté la régularité de ce phénomène, quelle que soit la taille des expressions fautives.

Tout d'abord, nous avons repéré une expression isolée *vous dorlotez avec son hospitalité*, pour le thème *personnel serviable* de la catégorie *description de l'hôtel*. En cherchant à la prendre en compte dans le cadre de la solution du lexique *traduction automatique,* nous avons réalisé que nous pouvions utiliser la stratégie des éléments minimaux pour réduire cette expression au simple mot-clé *hospitalité.* Dans ce cas, inutile de retoucher la base de connaissances, le mot-clé existant déjà pour le thème *Personnel serviable.*

De la même manière, pour plusieurs expressions contenant une erreur de traduction automatique telles que *location de voitures bureau* et *voiture de location d'entreprise localisée sur place* nous notons la possibilité de détecter le thème sans modification de la base de connaissances. Avec cette paire d'expressions, si nous ne gardons que les éléments minimaux, nous obtenons une forme parfaitement compatible avec des descripteurs du thème *location de véhicules* (catégorie *services de l'hôtel*), parmi lesquels *location de voiture(s)* et *voiture(s) de location*.

Ce phénomène se vérifie aussi avec des expressions en plus grand nombre. Par exemple avec le thème *Agence de voyages/Bureau d'excursions/Vente de billets*, nous ne trouvons quasiment aucune expression traduite correctement : *excursions de marche, excursions et excursions, excursion/a guidé des réservations d'excursion, aide d'excursion*... Mais le seul mot-clé *excursion(s)* permet la détection du thème concerné pour toutes ces exceptions.

Cette découverte pourrait appuyer le choix de la solution de *traduction automatique*, mais celle-ci n'est pas parfaite : elle peut faire gagner du temps sur la détection de certains thèmes mais peut aussi faire perdre de l'information.

E - Perte d'information à cause de la traduction automatique

Si la création d'un lexique *traduction automatique* n'a pas posé problème jusque là, la perte d'information totale due à une erreur trop ambigüe de traduction en est un.

La traduction du terme *hiking*, pour le thème *activités sportives* donne *hausse*. Totalement isolé, le terme ne peut être enrichi pour être détecté comme activité sportive. Il ne peut pas constituer un générique pour le lexique *traduction automatique* puisqu'il peut être employé dans plusieurs contextes (*sans hausse des tarifs, faible hausse des prix*…). Le thème ne sera donc jamais détecté, nous perdons de l'information à cause d'une erreur de la part du système de traduction automatique.

Nous retrouvons le même cas avec une traduction encore plus approximative : *accès élogieux aussi bien que les ordinateurs de wifi* pour le thème *internet gratuit* de la catégorie *services de l'hôtel*. L'expression *accès élogieux* est déjà transformée par le lexique TA comme signifiant *gratuit;* mais pour le reste, il semble plutôt difficile de créer une équivalence au générique *internet gratuit* avec une telle phrase.

Néanmoins, les cas de ce type sont rares ; il est toujours possible de détecter un ou plusieurs thèmes en prenant en compte les éléments minimaux d'une expression.

Chapitre 3 – Remarques importantes

La conception du lexique pour la traduction automatique nous permet de relever une caractéristique de ce lexique et d'identifier d'éventuels problèmes pour l'évaluation des solutions.

A - Le lexique TA : un lexique spécialisé

Au fur et à mesure de l'élaboration du lexique TA, nous avons noté la nécessité de créer des génériques très spécifiques.

Par exemple, les génériques *bar à vin* et *mini bar,* qui sont pourtant proches de sens, doivent être séparés puisqu'ils renvoient chacun à un thème différent.

Il est parfois nécessaire de séparer les génériques d'équivalences faisant partie du même thème, comme *kite_surf* et *activités sportives*. Le nombre d'équivalences oblige parfois à les cloisonner, la lisibilité et le retour au lexique sont plus aisés. Il est sans cesse nécessaire de revenir sur un lexique déjà créé ou sur la syntaxe d'une équivalence.

Si à chaque descripteur du thème correspond un seul générique du lexique TA, il est plus facile de le retrouver et de le corriger. A l'inverse, regrouper des équivalences différentes même si elles viennent d'un même thème, complique énormément la tâche lorsqu'il s'agit d'opérer des modifications.

Les génériques du lexique TA sont plutôt orientés vers un thème ciblé; nous trouvons peu de généralistes, tournés vers des concepts comme hôtel, enfant, payant ou disponible.

B - Évaluation : Difficultés de comptage par thématique

Pour l'évaluation, nous avons pris note du temps de création des génériques pour chaque thème. Il nous semble que l'objectivité de tels relevés est à remettre en question.

Nous venons d'expliquer les difficultés causées par les nécessaires et fréquents retours sur des éléments déjà créés. Le parallèle peut être fait sur le calcul des durées de mise en place, il est assez complexe de mesurer le temps de création des génériques d'un thème, ou même d'une catégorie, notamment de par l'imbrication (cf. chapitre précédent, partie B).

Il faudra en tenir compte lors du montage de l'architecture d'évaluation.

La finalisation de la mise en œuvre des deux solutions de traduction va nous permettre de les comparer et de les évaluer.

Partie 4

Solutions de Traduction : Évaluation

Chapitre 1 – Architecture d'évaluation

Nous venons de voir la mise en œuvre des deux solutions de traduction, il s'agit maintenant d'examiner comment il est possible de les évaluer.

A - Évaluation par thèmes

Afin de disposer de données temporelles pour l'évaluation, notre première idée était de calculer le temps de création des descripteurs ou des équivalences par thème. Nous aurions pu ensuite comparer leurs durées de mise en place selon la solution. Cette façon de procéder n'est pourtant absolument pas évaluable.

Nous nous sommes rendus compte que la mise en œuvre d'un thème ne se résume pas à la création de ses descripteurs. Quelle que soit la solution de traduction (une base de connaissances par langue ou une base multilingue), nous faisions de fréquents retours sur des thèmes déjà traduits ou des lexiques déjà créés. Au fur et à mesure que nous avancions dans le corpus, nous découvrions des expressions dont la forme pouvait inspirer un ajout ou une suppression d'un ou plusieurs éléments d'un descripteur ou d'un lexique. Dès le début du projet, nous avons compris que la comparaison du temps de création des thèmes n'était pas un bon indicateur.

Nous nous sommes donc tournés vers d'autres points de comparaison possibles.

B - Évaluation par étapes méthodologiques

Au lieu de concerner les thématiques, l'évaluation de la durée de la mise en place des deux solutions de traduction pourrait viser les étapes principales de la méthodologie.

1) Données comparables

Nous avons compté le temps de réalisation de chacune des étapes, il nous suffirait de comparer ces durées pour procéder à une évaluation. Auparavant, il nous faut savoir quelles données seront comparées et si elles sont comparables.

Reprenons les phases nécessaires à la mise en place de chaque solution.

En ce qui concerne la première solution - une base de connaissances par langue -, il a d'abord fallu constituer un corpus, c'est-à-dire choisir des descriptions d'hôtels et les mettre dans une base de données. Ensuite, en cherchant à obtenir la meilleure façon de traduire, nous avons créé un tableau de correspondances avec, par thèmes, le descripteur en français et sa traduction en anglais. Enfin, les traductions du tableau ont été reportés dans une base de connaissances et le reste du corpus a été transcrit directement dans celle-ci.

La deuxième solution - base de connaissances multilingue - a nécessité plus d'étapes. Premièrement, nous avons traduit dans la base de données le corpus anglais vers le français avec le système Systran. Il n'y a pas eu de phase de choix de fiches hôtel, nous avons seulement constitué le corpus par traduction d'un existant. Puis, une interface de lecture des fiches hôtel et de leurs traductions a été créée et nous avons aussi dû mettre à jour la base de connaissances en français. Pour finir, le lexique TA a été produit, en passant par un tableau de correspondance pour la première moitié du corpus et directement dans la base de connaissances pour la suite.

Le tableau ci-dessous récapitule les temps de réalisation de toutes ces étapes. Les durées sont calculées sur une base horaire, les chiffres du tableau de comparaison correspondent aux heures de travail effectuées.

Étapes de réalisation	Solution 1	Solution 2
Constitution Corpus	14	0
Création Base de données	14	10,5
Réalisation Interface	0	24,5
Tableau de correspondances	17,5	19,6
Mise à jour BdC française	0	94,5
Conception BdC	52,5	0
Élaboration lexique TA	0	73,5

Lexique des abréviations : *BdC* correspond à *base de connaissances* et *TA* à *traduction automatique.*

Fig. 10 : Tableau de comparaison des durées de réalisations

Ce tableau démontre que l'on peut mesurer les durées de réalisation au niveau de la méthodologie mais il n'est pas encore un outil d'évaluation. En effet, toutes ces étapes ne peuvent être comparées entre elles : certaines ne doivent être comptabilisées.

2) Données classables

Nous allons tenter d'ajuster les données afin de rendre possible une évaluation de leur contenu. Avec le tableau de comparaison, les points communs sont mis en valeur mais les différences d'étapes méthodologiques sont aussi visibles.

La solution de la base de connaissances unique et multilingue a nécessité deux phases supplémentaires par rapport à la première solution : la création d'une interface de lecture ainsi que la mise à jour de la base de connaissances française. Cependant, elles ne doivent pas être prises en compte dans l'évaluation. En effet, si la solution avec une base de connaissances unique et multilingue doit être à nouveau mise en marche, l'outil d'interface et la base de connaissances française à jour seront réutilisés. Les deux phases constituent donc des étapes de construction d'outils d'appui à la mise en œuvre d'une solution et le temps imparti à leur création ne doit pas être compter.

Il faut rappeler que l'étape du choix des fiches hôtel n'a pas été nécessaire pour la seconde solution - celle utilisant le lexique TA -. Le corpus utilisé dans cette solution a simplement été traduit depuis un pré-existant. Cela ne biaise pas l'évaluation, pour l'étape du choix des fiches hôtel, la seconde solution aura simplement un temps de réalisation égal à zéro heures de travail.

Ce sont les seules divergences de deux méthodologies, toutes les autres phases se révèlent superposables. Nous allons les organiser de façon pertinente et les classer en vue de pouvoir faire une évaluation.

Les réalisations des deux solutions ont en commun deux étapes principales, qui sont elles-même divisibles en deux parties comme suit :

- la constitution du corpus :
    - choix des fiches hôtel (valable uniquement pour la première solution),
      Cette phase consiste à rassembler des descriptions d'hôtels sur Internet, en prenant garde à diversifier le standing et la localisation des établissements.
    - création de la base de données,
      Les fiches hôtel choisies sont intégrées dans une base de données SQL.

- l'adaptation de la base de connaissances :
    - création du tableau de correspondances,
      Pour pouvoir trouver la meilleure méthodologie, la première moitié du corpus est transcrite (traduction ou création d'équivalences) dans un tableur.
    - modification de la base de connaissances,
      Une fois la méthodologie validée, les données du tableur sont transcrites dans la base de connaissances et le reste du corpus est directement traité.

A présent que nous savons quelles données utiliser et lesquelles se correspondent, nous allons pouvoir les comparer les unes aux autres.

3) Données évaluables

Nous allons maintenant comparer le temps de réalisation des étapes classées à l'aide du tableau suivant (volumes horaires):

Étapes de réalisation	Solution 1	Solution 2
Constitution Corpus	14	0
Création Base de données	14	10,5
Tableau de correspondances	17,5	19,6
Conception BdC / lexique TA	52,5	73,5

Lexique des abréviations : *BdC* correspond à *base de connaissances* et *TA* à *traduction automatique.*

Fig. 11 : Tableau d'évaluation des durées de réalisations

Grâce au classement des étapes des méthodologies, nous pouvons calculer le temps de réalisation de chacune des deux solutions de traduction. C'est dans le chapitre suivant que nous allons tenter d'analyser les résultats obtenus.

Chapitre 2 – Résultats d'évaluation

Les données évaluables nous aident à mettre au même niveau les deux solutions pour faciliter la comparaison.

A - Durée de conception

En additionnant les durées de chaque tâche, nous remarquons que la mise en œuvre des deux solutions de traduction est, à quelques heures près, de même durée.

Étapes de réalisation	**Solution 1**	**Solution 2**
Constitution Corpus	14	x
Création Base de données	14	10,5
Tableau de correspondances	17,5	19,6
Conception BdC / lexique TA	52,5	73,5
Total	**98**	**103,6**

Lexique des abréviations : *BdC* correspond à *base de connaissances* et *TA* à *traduction automatique.*

Fig. 12 : Tableau d'évaluation et résultats

Les deux solutions de traduction étudiées ont pourtant une méthodologie assez différente.

Pour utiliser un corpus, la première oblige à faire un travail minutieux de recherche et de tri de fiches descriptives d'hôtels tandis que la seconde nécessite seulement la traduction d'un corpus pré-existant en français.

Même si toutes deux sont passées par l'étape supplémentaire du tableau de correspondances pour la première moitié du corpus, la finalité est divergente. Dans un cas, nous disposons d'une nouvelle base de connaissances et il faut donc tout reprendre pour l'adapter à une autre langue. Alors que dans le second cas, il s'agit d'un ajout d'équivalences à un lexique dans une base de connaissances qui est déjà créée.

Lorsque la première solution prend du temps en constitution de corpus, la seconde allonge la durée de mise en place au moment d'imaginer un lexique supplémentaire qui, s'ajoutant à une base pré-existante, doit en saisir toutes les nuances pour comprendre les liens entre descripteurs et lexiques.

Il faut noter que les temps de réalisation affichés sont dépendants du fait qu'il s'agissait d'essais sur la meilleure solution de traduction. Ces durées doivent pouvoir être réduites puisque des étapes comme l'utilisation du tableau de correspondances ne seront pas appelées dans une prochaine mise en place des (ou d'une) solution(s) de traduction.

La première conclusion est donc plutôt inattendue : que l'on utilise un traducteur humain pour disposer d'une base de connaissances par langue ou un système de traduction automatique pour n'avoir qu'une base de connaissances pour plusieurs langues, la création d'un service Semantia multilingue mettra quasiment le même temps.

Évidemment, d'autres indicateurs que la durée de conception permettent d'évaluer ces deux solutions de traduction.

B - Aspects financiers

L'évaluation ne doit pas seulement tenir compte du temps de conception, elle doit aussi considérer le coût de cette conception.

La solution proposant un système d'une base de connaissances par langue fait appel à une traduction humaine et crée une nouvelle base de connaissances dans une langue étrangère. Dans cette approche, il faut disposer d'un bagage linguistique suffisant pour saisir les éventuels contre-sens que pourraient entraîner les descripteurs. Lorsqu'il s'agit de traduire en anglais, comme dans le cadre de ce stage, il est possible de trouver une personne de l'entreprise capable de mettre au point une traduction de qualité. Mais si Semantia a besoin de posséder une base de connaissances dans une langue moins connue que l'anglais, il est certain que Semantia devra faire appel à un traducteur humain extérieur à l'entreprise. La plupart des traducteurs, qu'ils soient à leur compte ou dans une agence, facture une journée de déplacement entre 250 et 400 euros. De plus, il faut prévoir une éventuelle formation de la personne employée aux technologies Semantia.

Par opposition, la solution d'une seule base de connaissances multilingue ne nécessite pas d'intervenants humains extérieurs à Semantia. Mais elle aussi, engendre un coût direct, celui de l'achat d'une licence pour un système de traduction automatique. Nous pouvons estimer ce budget à 800 euros, dans une fourchette de 200 à 1000 euros, prix pratiqués en général pour des logiciels de traduction automatique. Mais une fois le logiciel ou le service acheté à une entreprise sélectionnée, il nous faut encore l'intégrer au moteur Semantia afin de traduire le message voulu, avant son passage dans la base de connaissances multilingue. Il nous est difficile de prévoir le temps que devrait prendre cette intégration, étant donné que nous ne connaissons pas les moyens de procéder à l'ajout dans le moteur d'un système pré-conçu de traduction automatique. Nous savons, tout de même, que cette opération d'intégration ajoutera du temps de travail aux durées fournies auparavant (partie A).

Les deux solutions nécessitent du temps pour l'adaptation du traducteur aux technologies Semantia : une formation pour un traducteur humain et une intégration au moteur pour un automate.

Mais en comparant avec le coût de l'achat d'une licence d'un système de traduction automatique, l'embauche même ponctuelle d'un traducteur paraît plus couteuse.

C – Maintenance Évolutive

Nous devons aussi réfléchir la pérennité de chaque solution de traduction. Pour ce qui est de la maintenance de ou des bases de connaissances multilingues, nous pouvons nous interroger sur la flexibilité et la réactivité des deux solutions : laquelle nécessite une mise à jour plus régulière.

En comparant avec des mises à jour faites sur des bases de connaissances de taille similaire, nous savons qu'une maintenance mensuelle de plus de 14 heures doit être envisageable. Il est raisonnable de penser que plus on avancera dans l'utilisation de cette méthode, moins les descripteurs auront besoin de modification. Mais cette durée est tout de même dépendante de la solution de traduction choisie.

Pour la première solution, celle qui propose la mise en place de plusieurs bases monolingues, nous risquons d'avoir besoin d'une maintenance beaucoup plus importante. Il faut, en effet, tenir compte de la multiplication des besoins de maintenance en fonction du nombre de base de connaissances créées. Avec cette solution, à chaque langue ajoutée aux services Semantia correspondra une base entière dans une langue étrangère à mettre à jour. Le temps nécessaire à la maintenance sera donc augmenté de plusieurs heures à chaque création de base de connaissances.

Il n'en est pas de même pour la solution utilisant la traduction automatique. La maintenance sera potentiellement moins longue. En effet, quelque soit la langue, les thèmes, les descripteurs et les lexiques autres que le lexique TA restent un dénominateur commun. La maintenance se fera plutôt sur les génériques du lexique TA, ce qui représente moins de traitement par rapport à la maintenance d'une base de connaissances entière. De plus, nous pouvons nous permettre de supposer que les erreurs des systèmes de traduction automatique seront répétitives. La maintenance devrait donc finir par être de moins en moins nécessaire au fil de l'utilisation de la base de connaissances multilingue, et cela malgré la spécialisation du lexique TA (partie 3, chapitre 3, A).

Utiliser la solution de traduction d'une base de connaissances par langue entraîne irrémédiablement une augmentation du besoin de maintenance, tandis qu'en faisant appel à la deuxième solution, nous réduirons (ou au moins, stabilisons) le temps nécessaire pour mettre à jour la base de connaissances multilingue.

D – Conclusion de l'évaluation

Nous venons de voir les différences d'utilisation pour chaque solution de traduction. Il faut également élargir nos réflexions sur les résultats à un contexte plus global, avant de donner notre conclusion.

En effet, même si dans le cadre du stage seules deux langues seront étudiées, le principe retenu sera vraisemblablement étendu à cinq langues au moins. Nous disposons d'un meilleur niveau en anglais que pour une autre langue étrangère, la traduction en a été nécessairement facilitée. Il faut tenir compte des différences d'accessibilité aux langues étrangères : pour une langue peu « commerciale », un traducteur humain sera peut-être plus pertinent qu'un traducteur automatique.

De plus, nos tests ne concernent qu'un domaine particulier, celui de l'hôtellerie. La solution de traduction qui sera retenue devra être valable pour tous les secteurs d'activités dans lesquels Semantia est susceptible d'intervenir. Or, nous savons qu'un traducteur automatique est meilleur lorsqu'il vise un domaine particulier et puisqu'il coûte visiblement moins cher, il serait peut-être une meilleure solution. Par contre, les équivalences constituant le lexique TA devraient varier selon le système de traduction automatique intégré au moteur. Ils ne font pas tous les mêmes erreurs au mêmes endroits. La seconde solution de traduction est donc absolument dépendante d'un système n'appartenant pas à Semantia, ce qui pourrait se révéler un inconvénient majeur.

Nous remarquons que la solution de traduction la plus appropriée change selon les besoins (linguistiques ou thématiques).

Le choix de bases de connaissances monolingues se révèle plus intéressant pour une langue moins « commerciale », même si l'expertise d'un traducteur humain coûte plus cher ; et la solution d'une base de connaissances multilingue est tout désignée pour le type d'utilisation que fait Semantia de ses bases, c'est-à-dire une application à un domaine précis.

L'argument qui pourrait augmenter l'intérêt pour telle ou telle solution est celui de la maintenance. Son étude n'est pas l'objet de ce stage, mais il faudrait disposer de données chiffrées sur la possibilité de réduction du temps de maintenance d'un système utilisant un traducteur automatique.

Puisqu'une visibilité sur le court terme (temps de conception) tend à conclure vers une différence peu notable entre les deux solutions, une vision à plus long terme (maintenance) nous éclairerait certainement plus sur la pertinence d'une des solutions.

Conclusion

1) Bilan sur l'objet du stage

L'architecture d'évaluation mise en place nous a permis de mettre côte à côte la première solution « une base de connaissances par langue » et la seconde solution « base de connaissances multilingue ». Nous avons vu les avantages et les défauts de chacune : la première est valable pour traduire dans une langue peu traitée par les systèmes de traduction automatique, mais elle est coûteuse en terme de prix - rémunération du traducteur - et en terme de temps - maintenance -. La seconde sera plus efficace sur une base de connaissances détectant les thématiques d'un domaine précis, ce qui est souvent le cas. Elle nécessitera moins de coûts, financiers (licence uniquement) et humains (maintenance moins riche).

Néanmoins, il faut relativiser ces résultats. Avant la mise en œuvre de la seconde solution, nous avons mis à jour la base de connaissances française. La solution de la base de connaissances multilingue et sa création de lexique TA ont pu être facilitées puisque nous disposions d'une meilleure connaissance de la bases de connaissances, des outils qu'elle utilise ou propose. Certains mécanismes et raccourcis pour la création d'une base de connaissances sont devenus plus intuitifs pour la seconde solution ; cela a pu accélérer la mise en place de la seconde solution. De plus, les équivalences constituant le lexique TA ont été crées pour le système de traduction automatique Systran. Il se pourrait que l'utilisation d'un autre système requiert un temps plus court ou plus long pour la création des génériques.

Le travail effectué au cours du stage peut être amélioré. D'une part, la base de connaissances française qui nous sert d'appui pourrait encore être perfectionnée. Par exemple, nous pourrions testé la suppression des déterminants dans le message avant qu'il ne soit transformé dans la base de connaissances. Nous gagnerions du temps sur la création des descripteurs ainsi que sur le traitement lui-même. D'autre part, il serait intéressant de tester les solutions de traduction avec une langue moins proche du français que l'anglais, plutôt avec une langue non indo-européenne.

2) Bilan sur le stage

Le stage en lui-même s'est parfaitement déroulé, il m'a permis de découvrir les méthodes de travail et les différents pôles techniques d'une entreprise travaillant dans le domaine du Web et du Traitement Automatique du Langage Naturel.

L'équipe linguistique m'a donné l'occasion de travailler sur des projets en cours chez Semantia ainsi que sur des montages de Proof Of Concept (ou POC). Grâce à ces moments de travaux hors du sujet de stage, j'ai découvert de quelle manière les bases de connaissances étaient appliquées aux entreprises clientes. J'ai compris la nécessité de certaines nuances des bases de connaissances, comme l'existence des lexiques. Être intégrée à ces projets m'a permis de voir comment les services proposés aux clients étaient conçus, j'ai pu observer les différents étapes de la mise en place d'un projet.

Le domaine du « text mining » m'a toujours plu et j'ai été heureuse de travailler sur des projets professionnels dans ce domaine. J'ai intégré l'équipe de Semantia au 1er septembre, je suis ravie de débuter ma vie professionnelle dans cette entreprise.

Bibliographie

Ben Youssef, A. (2008) – *Méthodes mixtes pour la Traduction Automatique Statistique*, mémoire de master 2 recherche, Université de Grenoble, 54 p.

Darbelnet, J. (1970) - *Traduction littérale ou traduction libre ?,* Meta : Journal des traducteurs / Meta: Translators' Journal, volume 15, numéro 2, p. 88-94.

Gawron-Zaborska, M. (2000) – *Le fantôme de la traduction littérale dans la traduction juridiqu*e, La traduction juridique Histoire, théorie(s) et pratique, Colloque international organisé par l'Ecole de traduction et interprétation de l'Université de Genève et l'Association suisse des traducteurs, terminologues et interprètes.

Habert B. (2000) - *Des corpus représentatifs : de quoi, pour quoi, comment ?*, Linguistique sur corpus. Études et réflexions, Presses Universitaires de Perpignan, p. 3-5.

Péry-Woodley, M.-P. (1995) - *Quels corpus pour quels traitements automatiques ?,* Traitements probabilistes et corpus, TAL, volume 36, numéros 1-2, p. 213-232.

Ruwet, N. (1964) - *Compte rendu : Georges Mounin, Les problèmes théoriques de la traduction*, L'homme, volume 4, numéro 2, p. 141-144.

Sinclair, J. (1996) - *Preliminary recommendations on Corpus Typology,* Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).

Softissimo – *Livre Blanc, Logiciels de traduction instantanée*.
Consulté sur : www.softissimo.com

Sowa, J.F. (1976) - *Conceptual Graphs for a Data Base Interface*, IBM Journal of Research and Development, volume 20, numéro 4, p. 336–357

Vidrequin, C. (2008) – *Constitution de base de connaissances à partir de données textuelles non structurées*, Thèse doctorale d'université : Université d'Avignon, Laboratoire d'informatique (EA 931), 144 p.

Woodsworth, J. (1988) - *Traducteurs et écrivains : vers une redéfinition de la traduction littéraire*, TTR : traduction, terminologie, rédaction, volume 1, numéro 1, p. 115-125.

Liste des figures

UNIVERSITE Stendhal GRENOBLE 3

Déclaration anti-plagiat

DECLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.