Tóth Andrea: Relations sémantiques en traduction automatique: degrés d'inclusion et limites d'application
Argumentum 7 (2011), 227-239
Debreceni Egyetemi Kiadó

Tanulmány

Tóth Andrea

Relations sémantiques en traduction automatique : degrés d'inclusion et limites d'application[1]

Abstract

Statistical machine translation (SMT) often fails when faced with the problem of recognition of semantic relations in context. This article aims to give an overview and to detect possible reasons of semantic incoherence in translations. We try to demonstrate what effect the consideration of linguistic and natural language processing (NLP) methods can have on the quality of translations. We try to argue in favour of the introduction of word sense disambiguation (WSD) methods in the word selection process**.**
Keywords: Statistical machine translation, recognition of semantic relations, treatment of polysemy, word sense disambiguation

1 Introduction

La traduction automatique statistique (TAS) est confrontée à des défis multiples en ce qui concerne la reconnaissance des relations sémantiques. Le présent article vise à exposer la problématique de l'incohérence sémantique en combinant une approche strictement linguistique avec celle du traitement automatique des langues naturelles (TALN). Nous nous proposons de démontrer l'impact possible de la prise en compte de la textualité et de la sémantique sur la qualité des résultats de la TAS. Nous insisterons en particulier sur la nécessité de l'application des algorithmes de désambiguïsation lexicale (DL) dans la sélection lexicale.

Les systèmes de traduction automatique se subdivisent en deux types : il peut s'agir soit de traduction à base de règles, nécessitant des dictionnaires électroniques spécialement conçus ainsi qu'un système de règles linguistiques algorithmisé, soit de traduction statistique, nécessitant une sorte de mémoire de traduction. Les deux approches sont confrontées à diverses difficultés (et en engendrent même) ; on peut préférer l'une à l'autre, mais (à l'heure actuelle) elles sont certainement loin de pouvoir produire des traductions impeccables (ou même acceptables, dans la plupart des cas).

Les difficultés essentielles rencontrées en traduction à base de règles sont liées, d'une part, à l'approche de la langue en tant que code univoque et, d'autre part, au coût de l'élaboration de tels systèmes en termes de travail humain. C'est justement ce dernier facteur qui rend les

[1] L'auteur remercie István Csűry de ses très utiles suggestions, relectures et corrections.

solutions statistiques plus attrayantes. En effet, pour le succès de la traduction basée sur la statistique des équivalences interlinguales, il suffit de disposer d'une mémoire de traduction élargie. Or, cela ne va pas sans poser des problèmes car l'équivalent approprié d'un terme, d'une expression ou d'une structure grammaticale ne peut être obtenu qu'en cas de correspondance parfaite entre les items en les langues impliquées, enregistrés en tant qu'unités de traduction. Une combinaison des deux approches serait donc probablement plus efficace et, par conséquent, souhaitable.[2]

La raison pour laquelle on tend tout de même à préférer la TAS, c'est qu'il est relativement facile et économique de construire des mémoires de traduction, eu égard au grand nombre de textes (originaux et traduits) accessibles en ligne. Le coût peu élevé des ressources requises peut en quelque sorte récompenser la faible qualité des textes obtenus, par exemple, au niveau grammatical.

Malgré les possibilités qui restent actuellement très limitées en TAS concernant la prise en compte du niveau du texte, il n'est peut-être pas complètement inutile de considérer les leçons que la traduction automatique peut tirer des acquis de la linguistique textuelle. Il est clair que les textes produits par un système de TAS sont loin d'être parfaits, ne soit-ce qu'au niveau de la phrase. Ceci n'implique pas pour autant que l'on ne puisse adopter des approches textuelles pour l'amélioration de la cohérence et, d'une manière générale, de la qualité de ces traductions. La faible qualité des textes traduits automatiquement se ramène pour une large part à un facteur textuel important : la cohérence. Ces textes nous donnent une impression d'incongruité non seulement au niveau syntaxique, étant donné les phrases agrammaticales, mais aussi au niveau conceptuel ou sémantique, ce par quoi nous entendons ici l'usage inconséquent des sens, même ceux qui touchent au topic du texte. Le topic du texte est l'entité activée qui dénomme l'unité thématique la plus importante du texte par un substantif ou une référence anaphorique (Tolcsvai Nagy 2001). Les éléments relevant du topic du texte se répètent et leurs relations donnent une structure qui constitue le fondement de la cohésion lexicale. Pour que la cohésion lexicale propre au texte d'origine puisse subsister dans le texte traduit, les relations sémantiques diverses doivent être reconnues et reproduites en langue cible (comme la synonymie, la méronymie, l'hyponymie et antonymie des mots du texte) ou, d'une manière plus générale, les relations de coréférence (les relations anaphoriques), interprétées. Dans le cas d'un mot polysémique,[3] il est crucial que ce soit le sens adéquat au contexte auquel un équivalent en langue cible soit associé et que ce sens s'inscrive dans la chaîne de la cohésion du texte. Nous ne traitons de la cohésion lexicale que dans la mesure où les mots polysémiques y contribuent.

Le recours aux informations contextuelles est limité en TAS. En effet, les systèmes ne possédent pas de connaissances relatives au monde qui nous entoure (c'est-à-dire le contexte) ; ils ne peuvent s'appuyer que sur l'entourage verbal immédiat[4] (le cotexte) d'un mot polysémique afin de choisir l'équivalent adéquat en langue cible. En plus, ils ont du mal à gérer la complexité des relations qu'entretiennent les équivalents dans les deux langues.

---

[2] Prószéky (2006) présente un tel système, *MetaMorpho*, combinant les avantages des systèmes à base d'exemples (statistiques) et à base de règles.

[3] Le terme *polysémique* est employé ici dans le sens « mot ayant plusieurs sens », indépendamment du fait qu'il existe ou non une relation sémantique entre les différents sens. La dichotomie traditionnelle de l'homonymie et de la polysémie n'apparaîtra pas dans cet article.

[4] L'étendue du cotexte et le nombre des éléments pris en compte en tant qu' « entourage immédiat » diffère d'un système à l'autre et dépend largement des vues et des compétences des développeurs du logiciel en question.

Cette complexité relève des différences structurelles des champs lexicaux et sémantiques des différentes langues. Le cas échéant, il correspond à quelques-uns ou à chacun des sens différents d'un mot fortement polysémique d'une langue des formes différentes dans une autre, et vice versa.

Si l'on veut que ces relations sémantiques soient également prises en considération, il est évident qu'on doit recourir à d'autres méthodes que celle de la seule probabilité statistique des co-occurrences, appliquée en simple TAS. L'une des possibilités consiste à désambiguïser les mots à équivalents multiples, c'est-à-dire à déterminer lequel des sens possibles est activé dans un contexte donné. Nous traiterons de la désambiguïsation dans la section 5.

Évidemment, la question est celle de savoir si la tâche de développer un système qui tienne compte des distributions différentes des sens vaut l'effort en termes de performance.[5] En effet, la confection de dictionnaires électroniques (contextuels) ou l'annotation et l'étiquetage/ balisage de corpus parallèles constituent une tâche humaine énorme dont on ne peut qu'espérer que le résultat dépasse en qualité celui atteint par l'utilisation de l'algorithme le plus simple du choix de l'équivalent le plus fréquent. Il faut donc considérer si l'effort consacré est proportionnel au résultat attendu.

Dans ce qui suit, nous aborderons dans un premier temps un sujet linguistique en traitant des relations sémantiques qui tissent le texte. Ensuite, nous esquisserons brièvement ce en quoi consiste le choix, en TAS, des équivalents candidats en langue cible d'un mot qui est polysémique en langue source (sélection lexicale). Pour des fins d'illustration, nous utiliserons la traduction (du français vers l'anglais) de fragments courts, obtenue à l'aide d'un traducteur en ligne. Enfin, nous chercherons à démontrer l'intérêt de l'emploi de la DL en TAS par l'énumération des éléments qui pourraient potentiellement contribuer à l'amélioration de la qualité des traductions.

2 L'approche linguistique

L'étude de la cohésion lexicale porte sur les relations sémantiques intervenant entre les unités lexicales d'un texte. En suivant Halliday et Hasan (1976), il convient d'en distinguer deux types : la répétition et la collocation. (Les études ultérieures ne s'occupent que de la première catégorie, étant donné le manque de repères objectifs pour établir si tel phénomène relève ou non de la seconde.) Pour eux, la répétition n'est pas à prendre littéralement : ils y comprennent tous les cas de synonymie, d'hyponymie, d'antonymie et de méronymie ; leur interprétation est donc graduelle. Dans leur terminologie, la relation de cohésion impliquée par le terme collocation est déterminée moins par une relation sémantique intrinsèque que par une co-occurrence fréquente de leurs éléments dans des contextes similaires. Il ne s'agit ni des expressions idiomatiques ni des syntagmes ou locutions figées.

Pour expliquer le rapport entre la cohésion et la cohérence, Hasan (1984) introduit le terme de harmonie cohésive. Selon lui, le sentiment de cohérence résulte des relations sémantiques reliant les éléments lexicaux des textes.

---

[5] Et de rentabilité, bien sûr, ce qui est un autre problème, non moins important mais restant hors de la portée de la présente étude. Toutefois, il est intéressant de noter que si les méthodes statistiques de recherche (et de traduction) sont plus efficaces et plus rentables dans le cas d'un très grand nombre de requêtes des utilisateurs d'Internet portant sur un nombre restreint de sujets, elles se révèlent peu satisfaisantes dans le cas des requêtes « périphériques » sur des sujets extrêmement diversifiés, dont la part est bien loin d'être négligeable. (Communication orale de Ronald Kaplan (Microsoft) à Debrecen, le 31 mars 2011.)

Tóth Andrea: Relations sémantiques en traduction automatique: degrés d'inclusion et limites d'application
Argumentum 7 (2011), 227-239
Debreceni Egyetemi Kiadó

Il développe le rapport entre la cohérence et l'harmonie cohésive en élaborant la théorie des chaînes de cohésion. Dans cette théorie, il se distuingue deux types de chaînes : les chaînes correspondant, respectivement, aux relations d'identité et aux relations de similarité. Les relations d'identité établissent les liens de coréférence tandis que les relations de similarité gouvernent les liens de co-extension et co-classification. Parallèlement à cette distinction, une autre dichotomie comprend les relations instantanées, dont le rapport sémantique résulte d'une constellation concrète des éléments, et les relations constantes, dont le rapport est indépendant du contexte. Les études ultérieures n'examinent que les éléments de cohésion lexicale entrant dans ces chaînes (occurrences pertinentes) alors que les éléments non participant dans les chaînes (occurrences périphériques) ne sont pas pris en considération. Les éléments pertinents s'intègrent dans le sujet sémantique du texte et peuvent entrer en interaction entre eux. Les occurrences en interaction sont appelées occurrences centrales. Hasan définit le critère de cohérence d'un texte en établissant un seuil de valeur à atteindre : le taux des éléments centraux doit dépasser 50 % des éléments assurant la cohésion. Dans l'interprétation de Hasan, la cohérence n'est pas linéaire, comme l'organisation du texte ne l'est pas non plus.

Hoey (1991) essaie de donner une définition plus concrète de la relation entre la cohérence et la cohésion par l'analyse de la distribution des éléments assurant la cohésion lexicale. Selon lui, tout rapport cohésif relève de la catégorie de répétition mais se manifeste à divers degrés. L'étude de la cohésion lexicale doit se faire sur une base non linéaire car la cohésion relie non seulement les phrases qui se succèdent mais celles aussi qui se trouvent à des distances considérables à l'intérieur du texte ; lors de l'analyse de la cohésion, il faut donc considérer les relations parmi toutes les phrases du texte. Hoey considère deux phrases comme reliées s'il existe au moins trois relations de répétition entre eux. Les relations de répétition ont deux catégories principales: répétition lexicale et paraphrase, chacune ayant deux sous-catégories de relation (simple et complexe) qui se complète par la répétition de substitution entre des éléments non lexicaux (pronoms, adjectifs).

Les analyses de cohésion lexicale ont rendu possible d'établir d'une manière assez objective quelles sont les phrases centrales d'un texte et d'en faire ainsi un résumé à l'aide de ses phrases-clés. L'application des connaissances accumulées en linguistique (textuelle) au domaine de la traduction automatique pourrait contribuer à la résolution des problèmes comme celui de la désambigüisation des items lexicaux polysémiques. (Nous pensons à l'introduction des considérations sur les relations entre les occurrences pertinentes ou des collocations dans le sens de Halliday et Hasan comme domaines potentiels à intégrer.)

Après avoir évoqué les relations de nature essentiellement sémantique qui déterminent la cohérence d'un texte, voyons comment cette structure est traduite dans la TAS et quelles y sont les limites à ce que cette complexité de l'architecture puisse être reproduite. Nous passerons donc à l'étude du traitement de la polysémie en TAS.

3 Le traitement de la polysémie en TAS

Principes

Dans les logiciels de TAS, la désambiguïsation des mots polysémiques se fait uniquement sur la base des probabilités : le choix de l'équivalent est effectué par la sélection, parmi les équivalents potentiels d'un mot, de celui qui est le plus probable suivant une approche

statistique-combinatoire, sans la prise en compte de la contribution sémantique du mot au sens du texte.

En l'absence de capacité d'analyse sémantique et d'interprétation contextuelle, c'est sur une base statistique que l'équivalent du mot est choisi. Après avoir segmenté le texte de départ, le logiciel effectue une recherche dans une base de données bilingue contenant des textes parallèles et fait correspondre les traductions possibles au segment en question. Les traductions potentielles ainsi obtenues sont vérifiées sur une autre base de données qui contiennt des textes écrits en langue cible. Le logiciel choisit l'équivalent en fonction de la probabilité des combinaisons attestées de chaque élément du segment, le critère du choix étant le taux de probabilité le plus élevé. Malgré la chance élevée de tomber sur le sens adéquat, conforme au contexte à l'intérieur du segment, il est toujours possible qu'en ce qui concerne l'ensemble du texte, ce sens ne reflète pas les mêmes rapports que ceux qui constituent la cohésion lexicale du texte d'origine, c'est-à-dire que le mot qui est acceptable dans son environnement immédiat à l'intérieur du segment donné ne le soit pas par rapport au contexte de l'ensemble du texte. Le problème résulte du fait que la TAS n'est pas capable de prendre en considération le caractère non-linéaire de l'architecture des textes.

Un autre phénomène textuel peut également augmenter la complexité de la tâche de l'interprétation des relations sémantiques : le recours à des quasi-synonymes (synonymes contextuels) afin d'éviter les simples répétitions. C'est la recherche, par le locuteur/rédacteur du texte, de moyens d'alléger (ou d'agrémenter) son style qui est à l'origine de ce phénomène, responsable également de la multiplication des synonymes dans les textes de traduction. Károly (2007) décrit une tendance qui consiste à préférer les synonymes ou paraphrases à la répétition lexicale dans les textes de traduction (considérée comme l'un des universaux de la traduction).

Pour dépasser ces problèmes, il conviendrait de prendre en considération les relations sémantiques intervenant entre unités qui se trouvent les unes à une certaine distance des autres dans le texte. La gamme des types de relations est large : elle couvre les relations de répétition lexicale ou de collocation à partir de la synonymie en passant par l'hyponymie et l'homonymie jusqu'à l'antonymie.

3.1 La pratique du traitement de la polysémie en TAS

En guise d'illustration, prenons le mot (homonymique) *défense* et examinons son traitement dans les traductions de textes obtenues par le moyen du traducteur gratuit en ligne, Google Translate.[6]

Dans le texte suivant, les différentes occurrences du mot correspondent à ses valeurs sémantiques différentes. Il figure deux fois dans notre texte, dans des sens différents. Le système de traduction réussit tout de même à trouver le sens correspondant au contexte. Cette réussite vient du fait qu'ils apparaissent dans des contextes à co-occurrences récurrentes. La fréquence élevée des co-occurrences (*les défenses des éléphants, défense des animaux*) explique pourquoi le traducteur automatique est à l'aise pour trouver l'équivalent qui correspond au contexte actuel : si *défense* apparaît avec *éléphant*, un équivalent du sens «longue dent saillante d'animaux» est plus probable que tel autre signifiant « protection ».

---

[6] http://translate.google.com. Sur le fonctionnement du logiciel, voir les liens suivants:
Inside Google Translate (http://translate.google.com/about/)
Statistical Machine Translation (http://www.statmt.org)

Texte 1: Les braconniers chassent les éléphants pour leurs défenses. Suite aux protestations des organisations de défense des animaux, le commerce de l'ivoire est devenu interdit.[7]
Traduction automatique du texte 1 : The poachers hunt elephants for their tusks. Following protests from animal rights organizations, trade in ivory is now banned.

Dans le cas du fragment de texte qui suit,[8] le traducteur de Google n'est pas capable de reconnaître qu'il s'agit en l'occurrence du sens « interdire » et non pas celui de «plaider en faveur de, soutenir» de notre mot-cible, et cela malgré la présence même du mot *interdire* dans le contexte. La distance des deux mots explique que le logiciel ne tient pas compte des relations qu'ils entretiennent, c'est dans des segments non seulement différents mais aussi éloignés qu'ils se trouvent. Les mots extérieurs au segment dans lequel se trouve le mot cible ne peuvent pas servir d'indice, c'est uniquement l'environnement immédiat qui semble compter pour le système. Aussi l'effet du mot *solicitor* 'avocat' se prouve-t-il plus fort.

Texte 2 : En conclusion, j'interdis qu'aucune transaction soit essayée avec le déserteur Dufour. Je tiens d'avance pour sans valeur l'action de la justice anglaise dans une affaire intérieure de l'armée française (...). Je défends qu'aucun solicitor réponde pour moi ou pour mes subordonnés.

Traduction automatique du texte 2 : In conclusion, I forbid that no transaction is attempted with the deserter Dufour. I want you in advance for worthless action of English law in an internal affair of the French army (…) I advocate that no solicitor responds to me or my subordinates

[In conclusion, I forbid any transactions with the deserter Dufour. I consider, in advance, that any action in the British law courts in an affair which concerns the internal administration of the French Army (…) I forbid any solicitor to reply in my defence or that of my subordinates.]

Par les textes qui suivent, nous cherchons de démontrer combien le cotexte est important dans la sélection lexicale. Bien que le terme « point » figure dans quatre acceptions différentes dans le texte 3 (chacune relevant de domaines différents qui sont respectivement la géographie, la géométrie, l'expérience ordinaire et la mécanique), toutes ses traductions sont conformes au contexte, grâce à des éléments cotextuels fortement indicatifs : ils apparaissent dans des collocations (*les quatre points cardinaux*, *relier les points*, *point de départ*, *point mort*).

Texte 3 : Au plus fort de la tempête, le capitaine du navire contempla la carte sur laquelle figuraient les quatre points cardinaux. Il avait l'impression de tourner en rond. Il prit sa vieille règle en cuivre et relia les points qui indiquaient ses précédentes positions: la figure obtenue était un triangle. Il était revenu à son point de départ, comme si le moteur était resté au point mort.[9]

Traduction automatique du texte 3 : At the height of the storm, the ship's captain looked on the map that included the four cardinal points. He was going around in circles. He took his old rule copper and connect the dots that showed his previous positions: the figure obtained was a triangle. He had returned to its point of departure, as if the engine was stalled.

[At the height of the storm, the ship's captain looked on the map that included the four cardinal points. He felt like going around in circles. He took his old copper ruler and connected the dots that showed his previous positions: the figure obtained was a triangle. He returned to his point of departure, as if the engine had stopped in the dead centre.]

---

7 Ce « texte » est construit pour les fins de l'illustration.

8 L'exemple est extrait de l'article *défendre* du *Trésor de la Langue Française Informatisé* accesible à l'adresse http://atilf.atilf.fr/.

9 Le texte, trouvé à l'adresse http://cisad.adc.education.fr/eval/pages-98/telech/6e/les_pdf/ livretfreleve 986.pdf, provient des cahiers de l'évaluation à l'entrée en 6e, de septembre 1998 publiés par le Ministère français de l'Éducation nationale, de la recherche et de la technologie – Direction de la programmation et du développement (DP&D).

Tóth Andrea: Relations sémantiques en traduction automatique: degrés d'inclusion et limites d'application
Argumentum 7 (2011), 227-239
Debreceni Egyetemi Kiadó

Le texte 4,[10] qui ne contient pas d'indices facilitant l'interprétation du texte, est censé illustrer la difficulté qu'un logiciel (incapable de l'interprétation des relations sémantiques) doit dépasser lors de la traduction. Les humains qui construisent leurs interprétations à partir des informations tirées du cotexte aussi bien que de leurs connaissances personnelles, se rabattent sur ce qu'ils peuvent tirer de leur expérience dans la construction de la signification d'un tel texte, étant donné l'absence d'indices contextuels. Seulement, c'est au niveau de l'interprétation pragmatique que la traduction (ou l'interprétation) d'un tel texte pose des difficultés pour les humains, tandis que pour un logiciel de traduction, c'est au niveau sémantique que les difficultés apparaissent. Le terme polysémique *trou* est traduit en anglais par son équivalent *hole* (dont les sens sont acceptables dans chacun des contextes reconstruits par les sujets de recherche[11]), alors que le terme *prévisions* est une fois rendu par l'équivalent *expectations* 'attentes', l'autre fois par *forecasting* 'élaboration de prévisions (météorologiques)', ce qui déclenche des interprétations différentes.[12]

> Texte 4 : L'homme avait l'air préoccupé. Son front dégarni était barré d'une ride soucieuse. Derrière d'épaisses lunettes, ses yeux rougis clignaient sans cesse. Il continuait à mâchonner une cigarette éteinte qu'il ne songeait même pas à rallumer. Il se tourna vers son jeune collègue : – « Et ce trou » lui dit-il, « tu y as pensé ? Comment allons-nous faire pour le combler ? » L'autre avait l'air plus serein. La question ne semblait pas l'inquiéter et c'est presqu'en souriant qu'il répondit, d'une voix douce : – « Ne t'en fais pas, j'ai tout prévu ! Nous aurons tout ce qu'il faut demain matin et personne ne saura jamais qu'il y a eu un trou ! Toutes mes prévisions vont dans ce sens-là ! » et il ajouta, avec une pointe d'ironie : « tu sais que je fais entièrement confiance aux méthodes modernes de prévisions ».
>
> Traduction automatique du texte 4 : The man looked worried. His receding hairline was canceled by an anxious ride. Behind thick glasses, his eyes red blinking incessantly. He continued to chew off a cigarette that did not even think to rekindle. He turned to his younger colleague: – „And this hole" he said, „you've thought it? How will we do to address it?" The other looked more serene. The question does not seem to worry him and he said it almost in a smile, a gentle voice: – „Do not worry, I've thought of everything! We have everything you need tomorrow morning and nobody will ever know there was a hole! All my expectations going in that direction!" and he added with a touch of irony: „You know I have complete confidence in modern methods of forecasting."
>
> [The man looked worried. His receding hairline was creased with worry. Behind his thick glasses, his red eyes were incessantly blinking. He continued to chew his extinguished cigarette and did not even think to relight it. He turned to his younger colleague: – „And this hole" he said, „Have you thought about it? How will we do to fill it?" The other looked more serene. The question did not seem to worry him and he said, almost smiling, in a gentle voice: – „Don't worry, I've thought of everything! We have everything we need tomorrow morning and nobody will ever know there was a hole! I care about it all!" and he added with a touch of irony: „You know I have complete confidence in modern methods of foresight."]

Ces exemples nous permettent de constater qu'il est possible de rendre la signification correspondant au contexte en langue cible dans le cadre de la TAS dans les cas où le cotexte immédiat comporte des indices univoques, mais les tentatives risquent d'échouer en l'absence de tels indices. Or, nous avons pu également voir combien le choix de l'équivalent peut

---

[10] Le texte est volontairement flou, construit expressément à des fins de recherche, en matière de construction de la signification des textes polysémiques. Hollard, S. (2010) : Interprétation de textes polysémiques : une étude expérimentale appuyée sur l'oculométrie. *Glossa* 109 : 16-41. (http://www.glossa.fr/pdfs/109_ 20101022 114839.pdf)

[11] La recherche a montré que le texte ouvrait voie à 11 différentes interprétations (maçons, cambrioleurs, comptables, jardiniers, fossoyeurs, trésor, prisonniers, médecins, météorologues, policiers, autres) en ce qui concerne la profession des personnages.

[12] Sans parler de l'ambiguïté voulue par le locuteur (et signalée par la mention explicite, dans le texte, d'*une pointe d'ironie*).

Tóth Andrea: Relations sémantiques en traduction automatique: degrés d'inclusion et limites d'application
Argumentum 7 (2011), 227-239
Debreceni Egyetemi Kiadó

affecter la cohérence du texte. Les solutions à base statistique, si efficaces qu'elles soient à l'intérieur de segments plutôt courts, ne sont pas adaptées à reproduire les relations sémantiques au niveau du texte, affectant la cohérence textuelle. Nous pensons cependant que la cohérence est décisive pour l'équivalence fonctionnelle des traductions avec les textes de départ, si bien que, du point de vue de la signification de l'ensemble du texte, il est crucial que le sens (et l'équivalent) adéquat soit associé aux occurrences pertinentes, celles des mots et expressions établissant les réseaux de cohésion, facteurs clés de la cohérence textuelle. Or, pour y parvenir, il est indispensable d'inclure dans le système l'analyse des relations sémantiques, ce pour quoi les méthodes à base statistique sont assurément insuffisantes.

4 Le traitement de la coréférence

Dans une première étape, nous nous sommes proposé d'examiner le traitement des relations sémantiques dans le discours sur l'exemple de la polysémie, illustré par les textes 1 à 4. Dans une deuxième étape, nous évoquerons la problématique des relations de coréférence en traduction automatique, étant donné leur rôle essentiel dans la construction de la cohésion textuelle. Les problèmes de la reconnaissance des relations de coréférence étant fort complexes et dépassant les cadres de la présente étude, nous nous bornerons à les signaler à travers l'analyse d'un exemple.

Par relations de coréférence, il convient d'entendre les rapports d'identité référentielle (totale ou partielle) qui s'établissent entre les expressions référentielles d'un texte. Cela peut prendre diverses formes allant des simples répétitions lexicales aux paraphrases ou aux substitutions par des éléments sans autonomie référentielle tels que les pronoms ou les déterminants ou encore certains éléments de la morphologie verbale. Nous supposons qu'à part les problèmes plus spécifiques, comme l'interprétation de la référence des éléments contenant des noms propres, c'est plutôt ce dernier phénomène qui peut présenter des défis : la substitution d'un élément du texte par un élément grammatical, tel le cas de la pronominalisation. Nous allons donc analyser, dans ce qui suit, le traitement des pronoms d'un extrait littéraire en TA.

Texte 5 : « (1) En approchant de **son** usine, le père Sorel appela Julien de sa voix de stentor, personne ne répondit. (2) Il ne vit que ses fils aînés, espèces de géants qui, armés de lourdes haches, équarrissaient les troncs de sapin, qu'ils allaient porter à la scie. (3) Tout occupés à suivre exactement la marque noire tracée sur la pièce de bois, chaque coup de leur hache en séparait des copeaux énormes. (4) Ils n'entendirent pas la voix de leur père. (5) **Celui-ci** se dirigea vers le hangar en y entrant, il chercha vainement Julien à la place qu'il aurait dû occuper, à côté de la scie. (6) **Il l'**aperçut à cinq ou six pieds plus haut, à cheval sur l'une des pièces de la toiture. (7) Au lieu de surveiller attentivement l'action de tout le mécanisme, Julien lisait. (8) Rien n'était plus antipathique au vieux Sorel; il eût peut-être pardonné à Julien **sa** taille mince peu propre aux travaux de force, et si différente de celle de ses aînés; mais cette manie de lecture **lui** était odieuse, il ne savait pas lire lui-même. (9) Ce fut en vain qu'il appela Julien deux ou trois fois. (10) L'attention que le jeune homme donnait à son livre! bien plus que le bruit de la scie **l'**empêcha d'entendre la terrible voix de son père. (11) Enfin, malgré **son** âge, **celui-ci** sauta lestement sur l'arbre soumis à l'action de la scie, et de là sur la poutre transversale qui soutenait le toit. »[13]

Traduction automatique du texte 5 : (1) Approaching **its** plant, the father called Julien Sorel in his stentorian voice, no one answered. (2) He saw that his oldest son, a species of giant, armed with heavy

[13] Extrait du roman *Le Rouge et le Noir* de Stendhal, accessible sur le site du Projet Gutenberg : http://www.gutenberg.org/files/798/798-h/798-h.htm

axes, squared trunks of fir, they would bring to the saw. (3) Any busy to follow exactly the black mark drawn on the piece of wood, each stroke of their axes separated by huge chips. (4) They heard not the voice of their father. (5) **It** went to the hangar on entering, he sought unsuccessfully to replace Julien he should have occupied, near the saw. (6) **He** saw five or six feet high, straddling one of the pieces of the roof. (7) Rather than carefully monitor the entire mechanism of action, Julien read. (8) Nothing was more repugnant to old Sorel, he might have forgiven **her** slender little Julien appropriate to the work force, and so different from that of his elders, but this habit of reading **him** odious, he did not know read **it** himself. (9) In vain he called two or three times Julien. (10) The attention that the young man gave his book! well above the noise of the saw prevented to hear the terrible voice of his father. (11) Finally, despite **its** age, **it** jumped nimbly on the tree under the action of the saw, and there on the cross beam which supported the roof.

[„As he approached his mill, Pere Sorel called Julien in his stentorian voice; there was no answer. He saw only his two elder sons, young giants who, armed with heavy axes, were squaring the trunks of fir which they would afterwards carry to the saw. They were completely engrossed in keeping exactly to the black line traced on the piece of wood, from which each blow of the axe sent huge chips flying. They did not hear their father's voice. He made his way to the shed; as he entered it, he looked in vain for Julien in the place where he ought to have been standing, beside the saw. He caught sight of him five or six feet higher up, sitting astride upon one of the beams of the roof. Instead of paying careful attention to the action of the machinery, Julien was reading a book. Nothing could have been less to old Sorel's liking; he might perhaps have forgiven Julien his slender build, little adapted to hard work, and so different from that of his elder brothers; but this passion for reading he detested: he himself was unable to read. It was in vain that he called Julien two or three times. The attention the young man was paying to his book, far more than the noise of the saw, prevented him from hearing his father's terrifying voice. Finally, despite his years, the father sprang nimbly upon the trunk that was being cut by the saw, and from there on to the cross beam that held up the roof."][14]

Pour le lecteur humain, il est évident que ce passage est centré sur un référent, à savoir le père Sorel, représenté à la 3^{e} personne du singulier par les pronoms *il, lui, lui-même, celui-ci* ainsi que l'adjectif possessif (*son, sa, ses*). En même temps, trois autres personnages sont également présents, représentés eux aussi à la 3^{e} personne. L'un d'entre eux, Julien, reçoit une attention particulière si bien que les pronoms *il* et *le* (figurant sous la forme *l'*) et certaines occurrences de l'adjectif possessif (*sa, son*) se référant à lui font concurrence aux précédents qui renvoient à son père. La tâche de démêler ces réseaux de coréférence se révèle beaucoup plus compliquée pour le système de TA qu'une lecture humaine spontanée ne découvre de leur complexité. Nous ne mettons en relief que les traductions mal réussies des pronoms que nous avons indiquées par caractères gras dans les textes. (Les phrases des deux textes sont numérotées pour un repérage plus facile.)

L'erreur la plus saillante est sans doute le choix du pronom *it*, renvoyant normalement à un référent non animé, pour équivalent des substituts de *père Sorel*, brisant la cohérence aux niveaux local et global à la fois. Sa forme suffixée (*its*) correspond à *son* dans les phrases 1 et 11. Ce qui plus est, le pronom *it* semble être donné en équivalent même pour le pronom démonstratif *celui-ci* malgré la présence d'indices contextuels évidents et facilement accessibles (*leur père, son père* à la fin des phrases 4 et 10). Nous constatons donc que la frontière phrastique reste infranchissable aux algorithmes du système de TA qui, de toute façon, se heurtent à des difficultés élémentaires déjà sur le plan de l'identification des relations de coréférence intraphrastiques. Ceci suggère que dans la base de données des textes parallèles sur laquelle s'appuie le logiciel, *celui-ci* apparaît tout simplement en fonction de renvoi à des référents non-animés.

[14] La traduction est accessible sur le site du Project Gutenberg of Australia : http://gutenberg.net.au/ebooks03/0300261.txt

Tóth Andrea: Relations sémantiques en traduction automatique: degrés d'inclusion et limites d'application
Argumentum 7 (2011), 227-239
Debreceni Egyetemi Kiadó

À part la distinction fondamentale entre référent animé et référent non animé, les distinctions suivant le genre et/ou le sexe ainsi que la contrepartie sémantique des fonctions syntaxiques remplies par des pronoms, autant d'éléments cohésifs essentiels, permettent d'observer les faiblesses de la TA. Comme on le voit, le pronom *l'* est omis en anglais dans les phrases 6 et 10, les phrases traduites restent ainsi sans complément d'objet direct, ce qui constitue un obstacle sérieux à la compréhension au niveau local et détériore la cohésion de l'ensemble textuel. Dans la phrase 8, l'adjectif possesif *sa* renvoyant à Julien est traduit par *her,* ce qui montre que le système est incapable soit d'identifier le genre du nom propre *Julien*, soit d'identifier les relations de coréférence que les autres éléments du texte entretiennent avec lui. Toujours dans la phrase 8, le pronom personnel au cas datif *lui* est traduit par *him* en fonction de complément d'objet direct, tandis qu'un complément d'objet direct est ajouté dans la dernière proposition là où le texte de départ n'en contient aucun. Cela entraîne des distorsions au niveau syntaxique qui ont des répercussions évidentes sur le sens.

5 La désambiguïsation lexicale

La désambiguïsation consiste à déterminer la valeur lexicale à attribuer à tel mot (qui en a plusieurs) dans tel contexte. Pour ce faire, on doit disposer d'un inventaire des sens préalablement défini. Il existe différentes techniques pour la désambiguïsation : les unes appliquent soit des définitions extraites de dictionnaires, soit des ontologies rassemblant des informations non-linguistiques : les connaissances relatives au monde ; les autres utilisent des corpus d'entraînement sémantiquement annotés, à partir desquels des algorithmes (non ou partiellement) supervisés désambiguïsent l'ensemble des mots (*all words task*) ou un groupe prédéfini (*lexical sample task*). Dans le cas de la première approche, dite approche profonde (*deep approach)*, on obtient de meilleurs résultats, de qualité supérieure à ceux qui résultent de l'application du soi-disant approche superficielle (*shallow approach).* Pour définir de quel sens il s'agit dans un texte donné, on peut recourir à des définitions de dictionnaires : cette méthode est fondée sur le principe que les mots d'un texte sont sémantiquement liés et que cette relation sémantique s'observe dans la définition même des mots. Plus cette relation est manifeste (plus le nombre des éléments figurant dans la définition de chacune des valeurs du mot à désambiguïser est élevé et plus il s'en trouve dans son contexte actuel), plus il devient évident quel sens choisir. La plus grande efficacité pourrait être atteinte en matière de désambiguïsation par l'implication des ontologies mais il en existe une quantité très restreinte qui soit lisible par les ordinateurs si bien que cela ne fonctionne pas trop bien dans la réalité.

En ce qui concerne l'autre approche, celle des méthodes de désambiguïsation entièrement ou partiellement supervisées, l'optimisation de l'efficacité des algorithmes exige un travail humain d'annotation trop volumineux. Ce problème de collecte d'informations est bien caractérisé par la dénomination *goulot d'étranglement dans l'acquisition de connaissances* (*knowledge acquisition bottleneck* en anglais). Il pourrait être résolu par la méthode de désambiguïsation non supervisée dont le fonctionnement est assuré par la classification soit de contextes, soit de mots qui sont sémantiquement liés au mot cible. Le principe suivi est fondé sur l'observation que c'est dans des contextes similaires que les sens similaires apparaissent, d'où vient l'idée qu'un classement par contextes ou par mots est possible lors de la désambiguïsation (Schütze 1998).

N'ayant pas l'intention d'analyser, d'évaluer ou de comparer les différents algorithmes, nous nous contentons de donner un bref aperçu des techniques spécifiques de la DL

concernant la traduction automatique, sans avoir l'objectif de présenter les avantages et les inconvénients de l'un ou de l'autre.

Voyons ce qui remplit, en TAS, la fonction de désambiguïsation pour le traitement des mots polysémiques. Quelles sont, tout d'abord, les caractéristiques des bases de données lexicales, c'est-à-dire des inventaires de sens utilisés ? La plupart des outils de DL utilisent WordNet[15] (ou des systèmes apparentés dans les autres langues) en tant qu'ontologie lexicale. Cependant, il est crucial de choisir une base de données adaptée à la tâche à résoudre, qui est, en l'occurrence, la traduction automatique. Notamment, il faudrait par exemple prendre en considération l'écart entre les distributions des sens des équivalents dans le cas d'une paire de langues donnée et personnaliser la base de données en conséquence. Cela signifie qu'il faut supprimer les distinctions pour telle langue qui n'y sont pas pertinentes (dans les cas où la distribution des sens des équivalents coïncide) et, vice versa, en préciser de nouvelles là où cela s'avère nécessaire. Si on remplace cette base de données par la liste des équivalents repérés grâce au corpus d'entraînement du système de TAS utilisé, on arrive à éliminer les distinctions sémantiques non-pertinentes et d'en introduire d'autres qui le sont. Cette solution paraît faisable en raison du nombre relativement élevé des corpus parallèles alignés d'où on peut extraire les sens et leurs équivalents. Toutefois, l'emploi de ces équivalents et par cela l'établissement des correspondances bi-univoques entre sens et équivalents présente toujours un certain nombre de problèmes, puisqu'une approche qui présuppose que chaque équivalent véhicule un sens distinct unique ne tient pas compte de la complexité des relations entre les mots d'une langue et leurs équivalents dans une autre. Il serait pourtant nécessaire d'introduire une sorte de filtrage, de sensibiliser en quelque sorte les algorithmes à la détection des différences sémantiques puisqu'il est important de savoir si elles ne sont que des nuances ou, au contraire, si elles sont essentielles.

Pour résoudre le problème, Apidianaki (2009) propose une méthode de classement de sens. A l'aide de cette méthode de classement sémantique, un inventaire est établi de façon à ce que le regroupement des équivalents en classes reflète la distribution sémantique du mot ambigu. De cette façon, les équivalents sémantiquement proches sont distingués de ceux dont le sens est différent. Ainsi, il ne faut renoncer ni à l'inclusion des relations sémantiques ni à l'assimilation de la sélection lexicale à la DL (puisque la sélection de la traduction intervient en même temps que la désambiguïsation lorsqu'on fait appel aux équivalents de traduction au lieu d'un inventaire d'acceptions unilingue.)

6 Traduction automatique vs traduction assistée par l'ordinateur

Mais quels aboutissements peut-on espérer des efforts d'amélioration en matière de TAS ou, d'une manière plus générale, de TA ? Même si les développeurs réussissent à perfectionner les logiciels de TA par des algorithmes de plus en plus performants, notamment pour mieux faire face aux difficultés constatées en matière de sémantique, traduire sans intervention humaine semble être irréel, au moins dans un avenir prévisible : les constats généraux de Arnold et al. (1993) restent toujours valables.

[15] WordNet est une base de données lexicale pour la langue anglaise, dans laquelle les noms, les verbes, les adjectifs et les adverbes sont regroupés en ensembles de synonymes cognitifs (synsets), chacun exprimant un concept distinct. Les synsets sont reliés entre eux par des relations conceptuelles, sémantiques et lexicales. Wordnet est accessible par le lien http://wordnet.princeton.edu/.

Ceci ne revient pourtant pas à dire que ces applications soient indignes d'attention. Elles la méritent toujours mais doivent être traitées à leur place. A l'heure actuelle, c'est la TAO (traduction assistée par ordinateur) qui peut bénéficier davantage du développement de certains automatismes. L'emploi de la TAO permet d'économiser du temps et facilite la tâche des traducteurs à condition que la traduction de base soit de qualité acceptable. Les investissements en TALN peuvent être récompensés par une efficacité et une rentabilité accrues puisque la qualité des traductions faites automatiquement est en proportion inverse avec les efforts de correction du traducteur humain. Il est à noter que l'efficacité des automatismes atteint un point culminant dans le cas de domaines spécifiques (tandis qu'ils se révèlent peu efficaces dans les domaines génériques ayant un vocabulaire et des constructions diversifiés – la spécialisation permettant d'éliminer les significations hors de contexte).

Il existe encore d'autres domaines que celui de la traduction (automatique) où l'emploi des outils de TALN est bien justifié et se prouve rentable, donc efficace : par exemple, la facilitation de la compréhension des internautes confrontés à des textes écrits dans les langues qui leur sont inconnues[16].

Nous insistons ainsi sur la nécessité d'appliquer les acquis du TALN, tout en rappelant le fait non-négligeable : le recours aux pratiques de TA permet d'accéder pour un grand nombre d'utilisateurs à des informations autrement inaccessibles, en raison des limites de leur compétences linguistiques (absentes ou, du moins, insuffisantes). L'objectif de la traduction automatique (et par conséquence l'importance de son amélioration) est double : elle permet à la fois (1) d'assister les traducteurs humains en effectuant les travaux préparatoires ainsi que (2) de diffuser les informations accessibles en ligne à des masses d'utilisateurs et d'y accélérer l'accès.

7 Conclusion

Partant de l'assomption que les textes produits par les logiciels de TAS souffrent d'insuffisances considérables en matière de cohérence (à part leurs autres défaillances au moins aussi graves dans le domaine de la syntaxe, par exemple), nous avons examiné le potentiel de l'intégration de considérations sémantiques dans les applications de TAS.

Nous avons essayé de justifier l'intérêt de la combinaison des différentes approches relevant d'une part de la linguistique textuelle, d'autre part de la sémantique ainsi que d'énumérer les avantages de leur observation pour la performance de la TAS. Aussi avons-nous présenté des arguments en faveur de l'introduction des considérations sémantiques dans le procès de la sélection lexicale, en insistant sur l'importance de méthodes plus poussées de désambiguïsation lexicale.

Consciente de la complexité de la production de traductions sans contrôle humain et de l'impossibilité (au moins temporaire) pour les systèmes de TA d'égaler la qualité des traductions humaines, nous n'avons aucune intention de suggérer qu'une intégration plus poussée des considérations sémantiques puisse aboutir au remplacement des traducteurs humains par des logiciels de TA. Nous nous sommes contentée de décrire le traitement des relations sémantiques en TAS. Bien qu'il soit probable que la TA « parfaite », même avec des

[16] Un exemple éclatant de ce type est celui de MobiMousePlus. Prószéky et Földes (2006) ont développé un outil de compréhension sensible au contexte dans l'esprit de fournir appui aux internautes dans la traduction ou plutôt « décryptage » de textes en langues étrangères. L'outil fonctionne comme une sorte d'assistant à la compréhension, supporté par un certain nombre de dictionnaires électroniques.

solutions d'interprétation sémantique intégrées, reste du domaine de la science-fiction plutôt qu'une réalité, les recherches en vue de son amélioration sont plus que jamais actuelles. Notre ambition était de relever certaines de leurs faiblesses, observables au niveau de l'unité « naturelle » de la communication humaine : le texte.

Références

Apadianaki, M. (2009): *La place de la désambiguisation lexicale dans la Traduction Automatique Statistique* (Article présenté au 16e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009). Senlis, 24-26 juin 2009. http://www-lipn.univ-paris13.fr/taln09/pdf/TALN_73.pdf).

Arnold, D.J., Balkan, L., Meijer, S., Humphreys, R.L. & Sadler, L. (1993): *Machine Translation: an Introductory Guide.* London: Blackwells-NCC. (http://www.essex.ac.uk/linguistics/clmt/MTbook/)

Halliday, M.A.K. & Hasan, R. (1976): *Cohesion in English*. London: Longman.

Hasan, R. (1984): Coherence and cohesive harmony. In: Flood, J. (ed.): *Understanding reading comprehension*. Delaware: International Reading Association, 181-219.

Hoey, M. (1991): *Patterns of Lexis in Text*. Oxford: OUP.

Károly, K. (2007): *Szövegtan és fordítás*. Budapest: Akadémiai Kiadó.

Prószéky, G. & Földes, A. (2006): Between understanding and translating: a context-sensitive comprehension tool. *Archives of control sciences*, Vol. 15 n° 4, 637-644.

Prószéky, G. (2005): Machine translation and the rule-to-rule hypothesis. In: Károly, K. & Fóris, Á. (eds.): *New trends in translation studies (In honour of Kinga Klaudy)*. Budapest: Akadémiai Kiadó, 207-218. (http://www.morphologic.hu/downloads/publications/pg/2006_klaudy-book_mmo_pg.pdf)

Schütze, H. (1998): Automatic word sense discrimination. *Computational Linguistics* 24, 97-125. (http://www.aclweb.org/anthology/J/J98/J98-1004.pdf)

Tolcsvai Nagy; G (2001): *A magyar nyelv szövegtana*. Budapest: Nemzeti Tankönyvkiadó.

Tóth Andrea
Université de Debrecen
Département de Français
H-4010 Debrecen
Pf. 33
toth_andrea@vipmail.hu