Extraction d'un vocabulaire bilingue: outils et méthodes

Deryle Lonsdale
Center for Machine Translation
Université Carnegie Mellon
lonz@cs.cmu.edu

1. Résumé

Cet article présente les efforts liés à l'élaboration d'un vocabulaire bilingue pour un système de traduction. Nous présentons la méthodologie suivie pour identifier, normaliser et mettre en correspondance les vocabulaires anglais et français, puisés dans un corpus de textes préalablement traduits: analyse du texte source, compilation d'un index (mots simples et syntagmes), identification et traitement des composés nominaux, intégration des ressources terminologiques du client, alignement automatique des textes source et cible, extraction des traductions du vocabulaire source, rédaction du vocabulaire et son recodage en forme de données lexicales. Nous décrivons aussi les outils que nous avons conçus et développés pour nous aider à réaliser ces fins: indexeurs, éditeurs de contextes unilingues et bilingues, parseurs, programmes de décomposition des composés nominaux, et un éditeur de traductions.

2. Introduction

Parmi les différentes approches à la traduction automatique ou automatisée se classe le modèle dit "à base de connaissances". Certains essais préliminaires, comme par exemple le projet KBMT-89 [Goodman and Nirenburg, 1991] et le système apparenté KANT [Nyberg and Mitamura, 1992], utilisent, pour atteindre une traduction de haute qualité et un traitement automatique, une analyse sémantique relativement poussée. Cette analyse se base sur une représentation explicite des concepts et des relations qui les lient.

Depuis longtemps déjà, on constate qu'il existe une différence appréciable entre la création d'un prototype expérimental d'une part et l'élaboration d'un système de TA ou TAO utile et convivial d'autre part. Dans le cas des systèmes de traduction basés sur les connaissances, la commercialisation exige une saisie massive de données, qui ne peut se faire qu'avec une collaboration étroite entre client et fournisseur, facilitée par une technologie avancée. D'après [Galinski, 1988], le problème essentiel auquel sont confrontés les chercheurs dans ce domaine consiste en effet à obtenir des spécialistes techniques des données complètes et fiables pour chaque concept du domaine, à garder ces données à jour et à les restructurer automatiquement.

Ces derniers temps, on discute beaucoup l'utilité des ressources textuelles en ligne et les moyens de les exploiter pour l'acquisition des connaissances. Même si tout ne peut se faire

automatiquement, il faut minimiser le temps de participation des experts humains, tout en facilitant autant que possible leur travail. Ainsi le coût de développement sera moindre et l'expert (en traduction, terminologie, ou un domaine technique) ne sera pas trop dépaysé par l'apport d'informatique.

Nous décrivons ici nos expériences dans une telle entreprise: l'élaboration des ressources lexicales requises pour réaliser un système capable de traduire en français une quantité importante de documentation technique.

Nous ne traiterons pas ici de l'acquisition des autres formes de connaissances requises par le système, par exemple:

- l'agencement du modèle hiérarchique des concepts du domaine
- les règles grammaticales de structure syntaxique des langues source et cible
- les règles de restriction et de composition de l'interpréteur sémantique
- les règles de mise en correspondance qui assurent la réalisation de concepts neutres en langue cible

Toutefois, les principes et outils utilisés dans l'acquisition des données lexicales nous ont également servis dans ces autres tâches.

3. Identification du vocabulaire source

Dans le cadre du projet KANT, nous développons un système de traduction capable de traiter des textes rédigés en anglais contrôlé. Il nous a donc fallu fixer le vocabulaire anglais source pour le domaine traité: la documentation technique pour les équipements lourds. Ainsi, notre collaboration avec le client cherchait au début à d finir l'ampleur de vocabulaire et la complexité syntaxique requises pour permettre aux auteurs de s'exprimer aussi clairement, mais simplement que possible.

D'abord, nous avons dû évaluer les types de documents publiés, leur degré d'adéquation aux projets, et leurs traits principaux. Heureusement, le client possédait déjà des renseignements utiles concernant ces questions, ce qui nous a permis d'éviter une analyse détaillée avec des techniques d'extraction d'informations afin d'obtenir une caractérisation du domaine source. Le domaine particulier que nous avons décidé de traiter comprend donc les manuels de fonctionnement et d'entretien, les listes de pièces détachées, et les revues d'entretien, et ce pour plusieurs produits différents.

Nous avons reçu du client un corpus de plusieurs centaines de fichiers, comprenant quelque 53 megabytes de textes préalablement publiés, qui représentent le domaine choisi. Ces textes contenaient une quantité importante de tableaux, dessins, figures, énumérations et diagrammes, avec plusieurs codes de mise en page et de typographie, provenant des différents systèmes utilisés. Nous avons donc développé un ensemble de programmes et outils pour standardiser le contenu linguistique des textes, ce qui nous a notamment permis de normaliser

l'espacement entre mots, phrases, paragraphes, et mesures, ainsi que la ponctuation, les caractères spéciaux (à 8 bits, etc.), l'emploi des lettres majuscules, et les codes les plus importants de composition. Ces outils consistent d'une part en des sous-programmes UNIX (AWK, LEX, SED, etc.), et d'autre part, en des programmes codés en LISP et C.

Nous avons ainsi obtenu un corpus de fichiers de textes anglais qui s'est révélé relativement restreint: les 7 millions de mots (environ) ne contiennent que 12,000 mots uniques, dont une quantité importante de numéros de pièces, d'acronymes, d'initiales, et de noms propres.

Nous avons ensuite développé un indexeur codé en C qui enregistre la position de tous les mots simples qui figurent dans le corpus. Cet index sert, comme nous le verrons bientôt, de ressource pour les outils de contextes unilingues et bilingues.

Lorsqu'il est associé à un étiqueteur de codes grammaticaux, cet indexeur est aussi capable d'identifier et d'indexer des structures lexicales, ce qui nous a permis de trouver et stocker les composés nominaux,[1] les groupes verbe+préposition, et d'autres collocations. Pour ce faire, nous nous sommes servis du corpus Brown, mais nous avons dû y apporter certaines modifications. Par exemple, plusieurs mots techniques ne figurent pas dans ce corpus de référence: *screed*, *zener*, etc. En revanche, plusieurs catégorisations générales ne sont pas appropriées dans ces domaines techniques, où, par exemple, *keep* et *will* ne sont plus des nominaux. Moyennant de telles modifications, ce processus a permis de récupérer plus de 110,000 occurrences de composés nominaux distincts (y compris les flexions morphologiques).

Nous avons ensuite procédé à l'élimination des formes jugées inutiles, comme les termes qui présentent des fautes de frappe ou une orthographe non standard et certaines abréviations:

- exhaust gasses = exhaust gases
- fuel sulphur content = fuel sulfur content
- 9 tooth dog clutch = nine-tooth dog clutch
- digital lcd display = digital liquid crystal display

La normalisation de l'espacement, surtout avec l'emploi du trait vertical et du trait d'union, a permis de mettre en relation plusieurs centaines de termes :

- air to air aftercooler = air-to-air aftercooler
- blowby/airflow instrument = blowby/air flow instrument
- one way clutch = one-way clutch
- infra-red light = infrared light
- air/fuel mixture = air-fuel mixture

L'identification de variantes dérivationnelles ou flexionnelles a également réduit sensiblement le nombre de termes:

[1] Par "composé nominal", nous entendons les collocations de nominaux qui forment la lexicalisation d'un concept du domaine, ce qui inclut les syntagmes lexicaux et souvent les syntagmes libres: voir [Bédard, 1986].

- vibratory motor = vibrator motor = vibration motor
- air condition system = air conditioner system = air conditioning system
- operator platform = operators' platform = operators platform = operator's platform

Plus importantes sont les formes obtenues par ellipse, surtout suite à la présence dans le contexte du référent en question : la forme pleine se présente au début, suivie d'une ou plusieurs variantes du même terme, que l'anglais contrôlé n'admet pas:

- alkyd paint = alkyd type paint
- rops/fops structure = rops/fops protective structure
- aec switch system problem = aec pressure switch system problem

Les composés synonymes se forment quelque fois avec les mêmes mots, mais dans différents ordres:

- stabilizer/dozer control main valve = stabilizer/dozer main control valve
- idle timer shutdown feature = idle shutdown timer feature

Bien sûr, l'indexeur a identifié au début plusieurs formes erronées, en raison de leur ambiguïté. Par exemple, l'ambiguïté anglaise nom/verbe menait à l'extraction de termes incorrects (mis ici en italique) dans les contextes suivant:

- Techniques develop as the *operator gains knowledge* of the truck.
- Apply soda solution to the battery until the cleaning action of the *soda stops*.

Nous avons essayé de faire ce tri entre termes utiles/inutiles le plus automatiquement possible, avec des algorithmes comme l'**edit-distance**. Pourtant, nous avons souvent dû faire appel au corpus ou même à un expert du domaine afin d'évaluer un terme suspect. Par exemple, l'outil KWIC (décrit ci-après) montre que, contrairement à ce que nous avions soupçonné, *bendix drive* est une pièce, tandis que *executive drive* n'est qu'une rue dans une adresse. Par contre, un expert compétent unilingue a dû être consulté pour juger d'autres cas, surtout concernant les synonymes et le sens des acronymes. Suite à la normalisation d'usage pour toutes ces formes, la neutralisation des formes variantes, et le rejet des formes fausses, nous avons aujourd'hui un vocabulaire de base de quelque 60,000 termes.

Notre prochaine tâche consistait à déceler la structure des composés nominaux. Nous avons utilisé une approche plutôt conservatrice, estimant que ces composés sont formés d'une manière strictement binaire et compositionnelle (voir [Levi, 1978]). Ainsi, à chaque niveau de composition, il n'est possible de combiner que deux sous-unités. Une sous-unité n'est admise que si elle se manifeste indépendamment ailleurs dans le corpus.

Parfois, une seule possibilité se présente, au dépens de toutes les autres possibilités:

(auxiliary ((fuel filter) (housing assembly)))

Pour calculer une préférence en cas de décompositions ambiguës, nous avons employé un score basé sur la fréquence d'occurrence des sous-unités, comme dans les deux exemples suivant:

```
((bypass valve displacement)
  ((((bypass valve) displacement) 98%)
   ((bypass (valve displacement)) 2%)))

((spray system water tank level)
  (((((spray system) (water tank)) level) 80%)
    ((((spray system) water) (tank level)) 12%)
    (((spray (system water)) (tank level)) 4%)
    ((spray ((system water) (tank level))) 4%)))
```

Parfois, le système ne produit aucune décomposition; ceci arrive souvent avec des formes idiomatiques, bahuvrihi, dvandva, et quand il y a ellipse de sous-unité. [Warren, 1978] note que c'est précisément ce genre de composé qui ne suit pas l'hypothèse de composition binaire.

Le corpus anglais nous a aussi servi de référence pour déterminer la stylistique et les conventions méta-linguistiques, ainsi que le registre et le niveau de vocabulaire. Ces constatations se sont faites manuellement.

En somme, le corpus source est apparu comme une ressource capitale dans notre effort d'identifier et classer le vocabulaire d'origine et de trouver les unités lexicales intéressantes.

4. Identification du vocabulaire cible

Nous avons dû ensuite identifier le vocabulaire français. Nous avons suivi à peu près la même démarche que pour le vocabulaire anglais, en nous servant des mêmes outils. Plus modeste, le corpus français ne représente que 10% du corpus anglais.

L'extraction des codes de mise en page s'est révélé un peu plus complexe, vu l'abondance de codes différents pour les lettres accentuées et le besoin de les normaliser. Une simple extension à nos outils d'analyse source a néanmoins suffi.

Nous avons ensuite compilé les index pour les mots et syntagmes cible, ayant préalablement amélioré l'indexeur pour tenir compte de la syntaxe des ces syntagmes. En effet, nous ne disposions pas au début de toutes les données nécessaires pour l'étiquetage du corpus français, mais nos outils lexicographiques nous ont aidés dans cette tâche. L'indexeur a trouvé quelque 19,000 mots simples et 160,000 composés nominaux dans le corpus français, qui compte quelque 1,750,000 mots.

Une certaine normalisation était aussi possible dans le vocabulaire français. Puisque les résultats de cet effort ressemblent à ceux de l'anglais, nous ne donnerons pas d'exemples.

Notre outil KWIC (KeyWord in Context: mot-clé et contexte, Figure 1) est conçu pour afficher le contexte de n'importe quel mot ou syntagme figurant dans les index des corpus source et cible. Il est suffisamment général pour traiter l'anglais, ainsi que les autres langues envisageables dans un avenir proche. L'outil permet de voir les occurrences séquentiellement ou dans des contextes triés (à gauche ou à droite), avec longueur de contexte variable. Puisque cet outil fonctionne sur une gamme de plate-formes logiciel/matériel, nous avons dû restreindre sa fonctionnalité d'affichage et d'interaction.

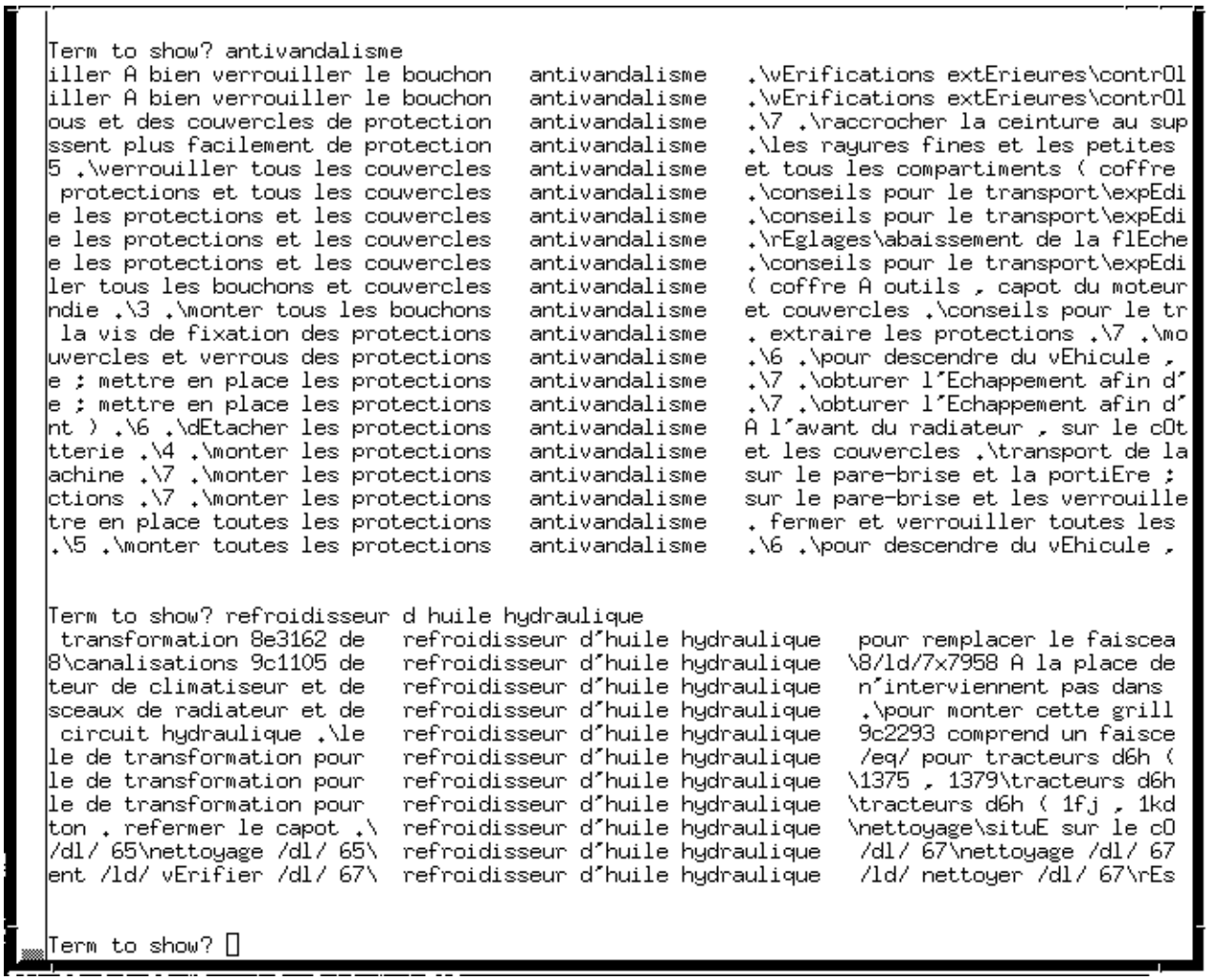

Figure 1: **Affichage de l'outil KWIC**

5. Appariement bilingue des vocabulaires

Puisque la collection largement automatique des vocabulaires était requise pour d'autres raisons, nous avons choisi d'associer autant que possible les unités lexicales des deux langues

suivant un processus d'appariement direct.

Pour commencer, nous avons réuni toutes les ressources lexicales bilingues du client. Celui-ci avait à sa disposition une banque de données bilingues utilisée pour remplir les commandes de pièces détachées. Comme c'est souvent le cas avec de telles ressources, le format des entrées ne convenait pas complètement; un remaniement avec des programmes LISP et une révision manuelle se sont imposés. Le client a aussi fourni un modeste lexique bilingue de termes simples, avec annotations, compilé par plusieurs traducteurs. Ce lexique, lui aussi, nécessitait une réorganisation presque totale. L'emploi de plusieurs dictionnaires bilingues en ligne nous a fourni quelques traductions pour chaque terme source général (simple et composé) . Nous avons rejeté automatiquement les traductions ne figurant jamais dans le corpus cible. L'intégration de ces trois ressources lexicales bilingues nous a servi de point de départ. Toutefois, de cette manière, nous n'avons pu récupérer que quelque 6,500 termes, soit 11% des termes source.

Nous avons ensuite entrepris une autre étape de collection d'équivalents source/cible, basée sur les deux corpus respectifs. Nous avons développé un programme d'alignement des corpus qui nous a permis d'accéder au contexte de tout terme source et de trouver le contexte correspondant du texte traduit dans un fichier de langue cible. A l'aide des index de termes source, nous avons trouvé chaque occurrence du terme et scanné les fichiers cible pour retrouver les contextes correspondants en français. Nous avons conçu et développé un outil sous le système d'interfaces X-Windows pour afficher les résultats de l'alignement. Il soutient aussi l'interaction avec un utilisateur, ce qui permet à ce dernier de sauvegarder dans une base de données les correspondances source/cible intéressantes.

L'alignement s'est révélé peu problématique, contrairement aux premiers soupçons. Malheureusement, en raison des divergences importantes et fréquentes entre les textes anglais et français, une approche basée sur les algorithmes statistiques à programmation dynamique (comme [Gale and Church, 1991]) n'a pas produit d'excellents résultats. Par contre, nous avons développé notre propre approche qui suit l'esprit de [Simard et al., 1992].

Nous avons tiré profit de certains éléments du texte qui restent invariables dans la traduction. Par exemple, le texte est particulièrement riche en diagrammes étiquetés qui ne changent jamais de langue en langue et dont les renvois restent invariables dans le texte. Il contient aussi une quantité importante de mesures, chiffres, et numéros de pièces. Chaque phrase est donc classifié d'après les points saillants les plus distinctifs de son contenu: mesures, renvois numériques, numérotisation, et catégorisation lexicale générale (ce dernier critère étant le plus lent et donc le moins usité). Une comparaison des codes de classification de phrases permet de faire défiler de façon parallèle les deux fichiers, combinant les phrases source ou cible au besoin.

Meme si la terminologie est assez difficile en raison de son caractère technique, nous avons trouvé que les personnes ayant une bonne connaissance générale de la langue peuvent au moins isoler et identifier ces unités terminologiques. Ainsi, nous avons employé plusieurs étudiants bilingues à temps partiel pour récolter ces termes dans les alignements.

Se servant de l'outil BiKWIC (Figure 2), ils prennent une liste de termes source à traduire

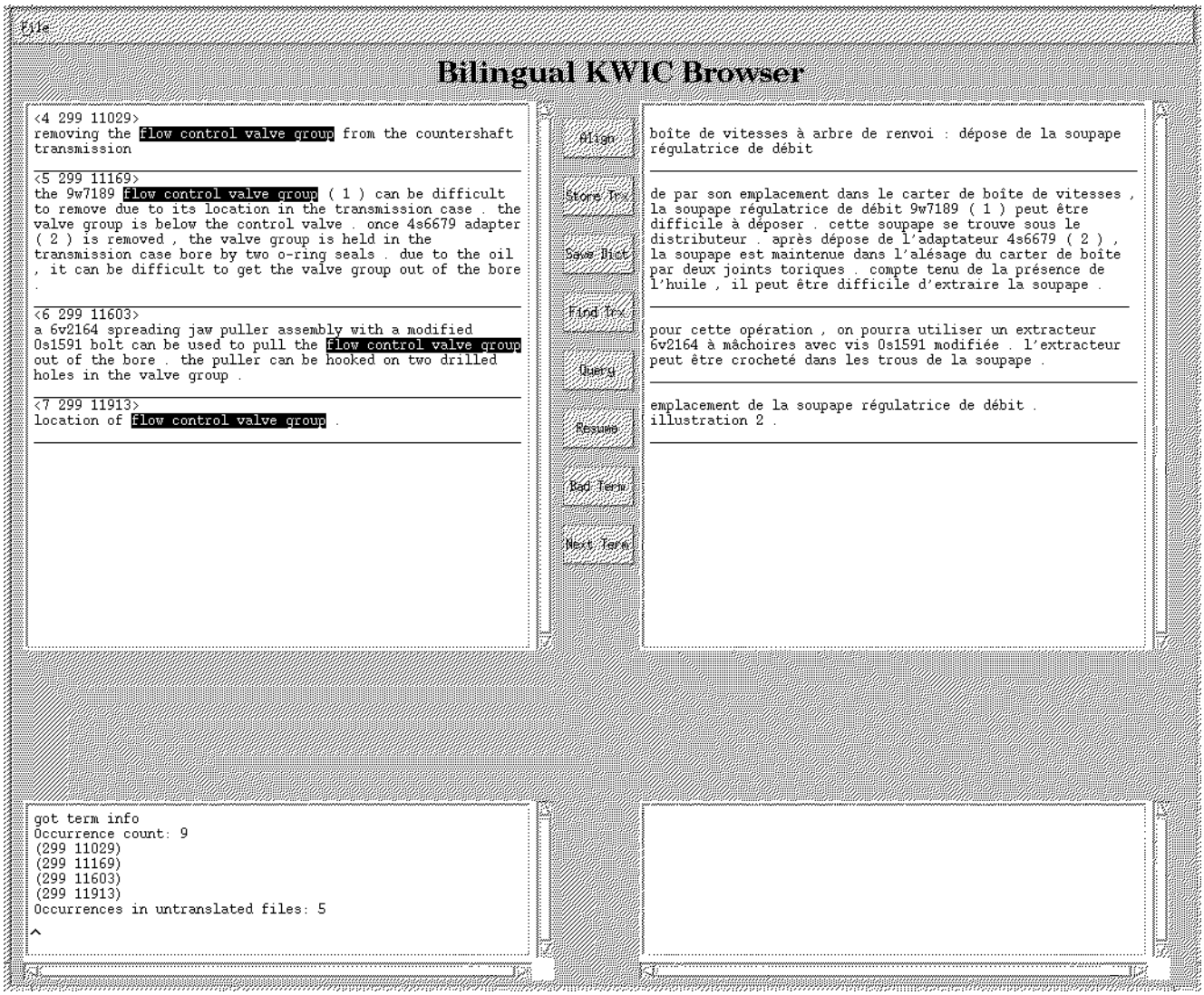

Figure 2: **Affichage de l'outil BiKWIC**

et examinent les contextes alignés dans les corpus source et cible. Lorsqu'ils trouvent dans ces contextes une correspondance traductionnelle, ils enregistrent la traduction avec une opération de souris dans une base de données terminologiques associée à l'outil, qui sera revue plus tard par un traducteur/terminologue expert dans le domaine. De cette façon, nous avons pu extraire les traductions de quelque 8000 composés nominaux et de 4000 mots simples. Un traitement plus sophistiqué d'alignement intra-phrastique nous permettrait d'extraire de façon automatique ces unités lexicales et leurs traductions; nous espérons faire davantage de preuves dans cette direction.

Toute cette démarche nous a donné un premier ensemble substantiel du vocabulaire, tiré des deux corpus reçus du client. Enfin, cette liste a été revue et les termes manquants fournis manuellement. Pour faciliter cette tâche énorme, nous avons conçu et construit un outil interactif pour l'enregistrement de ces termes. Un traductuer expert dans le domaine

s'est servi de l'outil pour faire ces deux tâches à la fois.

Cet outil de rédaction de vocabulaire affiche les termes anglais, un à la fois, avec tous les termes apparentés (c'est-à-dire qui ont au moins un mot en commun). Le cas échéant, les traductions du terme vedette sont aussi affichées. Pour chaque terme à traduire, le système essaie de proposer une traduction brouillon, en comparant celles des termes apparentés pour en tirer les généralités nécessaires à la construction par inférence d'une traduction. Les termes traduits les plus utiles sont ceux qui incluent le terme vedette comme sous-terme: un découpage de la traduction du terme entier s'impose dans ces cas.

6. Recodage en forme fonctionnelle

Une fois la correspondance lexicale établie entre les concepts neutres du domaine et les unités lexicales de la langue cible, il faut recoder celles-ci dans une forme plus fonctionnelle. Le générateur requiert une description de la structure syntaxique de ces unités; sinon, les flexions et autres modifications risquent de se faire incorrectement.

On peut considérer, à titre d'illustration, le problème de la représentation lexicale des composés nominaux français. Le sous-système exécuté pour générer le texte français de sortie suit un ensemble de règles grammaticales décrivant la morphologie et la syntaxe de la langue cible. Il opère sur des données que l'on appelle des "structures fonctionnelles grammaticales" (des SFG), les traitant selon un algorithme récursif et descendant. Chaque unité lexicale doit donc être recodée en SFG pour pouvoir être intégrée avec les autres constituants de la phrase. Ces SFG, qui peuvent d'ailleurs devenir plutôt complexes, peuvent elles aussi être créées automatiquement, et révisées au besoin.

La formation des SFG se fait à l'aide de notre parseur de texte français. Ce parseur comprend l'analyseur-compilateur généralisé gauche-à-droite qui convertit un ensemble de règles syntagmatiques de la grammaire française en une table d'analyse. Ces règles suivent de très près le format de celles du générateur. (Ces deux grammaires du français sont proches mais non pas identiques, ne supportant donc pas un traitement pleinement réversible analyse/génération.) Le parseur prend ainsi la forme lexicale du terme français pour le convertir en SFG.

Pendant cette conversion texte-SFG, nous avons rencontré les problèmes classiques: attachement des syntagmes prépositionnaux, adjectifs, etc., ce qui cause souvent la multiplication des représentations candidates. Par exemple, si le terme *front main control valve bank* se traduit *bloc de distributeur principal avant*, il faut résoudre l'attachement du mot *avant* pour représenter la structure correcte du terme français. Parmi les 10 compositions admises par le corpus, celle jugée la plus acceptable d'après nos métriques se forme ainsi:

(front ((main (control valve)) bank))

Admettant, d'après [Bauer, 1978], une généralisation parallèle en français, on obtient la SFG

suivante:

```
((CAT noun) (ROOT bloc) (POSTMOD -)
 (AGR
    ((NUMBER sg) (GENDER m)))
 (MODIFIER
    ((CAT adj) (ROOT avant)))
 (PP
    ((P-OBJ
        ((ROOT distributeur) (CAT noun)
         (AGR
            ((NUMBER sg) (GENDER m)))
         (MODIFIER
            ((CAT adj) (ROOT principal)))))
     (PREP
        ((ROOT de) (CAT prep))))))
```

Nous avons dû intervenir pour de tels attachements ambigus; heureusement, nous avons pu nous aider du corpus des décompositions pour l'anglais.

Ces unités lexicales et leurs formes SFG sont utilisées dans le système par le mappeur, qui assure la réalisation des concepts primitifs interlangues en structures linguistiques de la langue cible. Nous ne pouvons pas entamer ici une discussion de cette étape; il suffit de noter que la collection de toutes ces classes de données requises par le mappeur et le générateur constitue une partie essentielle du système.

7. Améliorations futures

Durant les travaux du projet discuté ici, nous avons dû mettre en balance les techniques automatiques et parfois expérimentales, et les approches reconnues mais moins informatisées, ce qui nous a parfois forcé à remettre à plus tard un traitement de pointe prometteur mais qui n'a pas fait ses preuves. Dans ce chapitre, nous mentionnons certaines des améliorations possibles.

Notre liste étiquetée de mots anglais ne réflétait pas complètement la nature technique du corpus source. Pour cette raison, les structures lexicales contenaient parfois des erreurs nécessitant une révision manuelle. Un meilleur ensemble d'étiquettes, ou une heuristique d'induction de catégories grammaticales pourrait nous aider à ne retirer que les unités lexicales qui nous intéressent.

Nous n'avons pas utilisé de techniques de troncature, comme celles traitées dans la littérature d'extraction (par exemple [Savoy, 1991]), pour neutraliser les flexions des formes morphologiques. Pour des raisons pragmatiques, nous avons développé des algorithmes et

programmes ad hoc; ils ont été utiles dans notre cas, mais ils ne se généralisent pas particulièrement bien. Nous comptons intégrer une approche plus satisfaisante.

Nous allons aussi améliorer les heuristiques d'identification des structures lexicales intéressantes. Par exemple, nos sous-programmes, écrits actuellement en C, peuvent être extraits avec l'indexeur et recodés en une langue telle LEX qui admet plus facilement une spécification de structures syntagmatiques. Nos outils seront alors plus facilement extensibles.

Notre technique d'alignement bilingue, bien que satisfaisante, dépend beaucoup du caractère spécial d'une sous-classe des textes du client. Si nous comptons, à l'avenir, traiter des textes sans mesures, renvois, etc., il nous faudra y intégrer les techniques récemment decrites, par exemple, dans [Church, 1993].

L'extraction d'équivalents ne s'est pas faite ici de façon totalement automatique, ce qui s'explique par notre manque de ressources lexicales primaires pour ce projet, mais aussi par le fait que les techniques d'appariement lexical sont toujours sous investigation. [Debili and Sammouda, 1992] présentent une solution éventuelle possible.

Nos premiers essais en description compositionnelle ont fourni une quantité sensible de données, mais nous n'avons pas encore trouvé la façon de profiter pleinement de cette ressource dans la spécification bilingue des unités lexicales. Cela nous semble être une démarche riche en possibilités.

Au cours des efforts de développement du vocabulaire, nous devions souvent consulter un expert dans le domaine approprié. Même des questions de synonymie, de polysémie, et d'usage nonstandard ont parfois nécessité ces interventions. Ceci peut être un processus lent et difficile, mais aussi coûteux. Nous outils KWIC et BiKWIC nous ont souvent aidé à résoudre certains de ces problèmes; toutefois, il est sans doute possible de relever davantage de renseignements directement des corpus par des moyens plus automatiques.

Finalement, il était souvent difficile d'arriver à des chiffres exacts lorsqu'il fallait quantifier nos efforts au préalable: temps de traduction en fonction de nombre de termes, grandeur des corpus et résultats de l'analyse, ampleur du vocabulaire final et des formes à rejeter, ressources informatiques nécessaires, etc. Le développement de meilleures techniques d'estimation et de quantification devrait servir à l'avenir à réduire le coût et les risques associés au développement de vastes bases de connaissance.

8. Conclusion

Nous avons mentionné ici un ensemble d'outils et de méthodes d'acquisition de la connaissance lexicale, mis au point pendant le développement d'un système de traduction à pivot interlangues. Sa mise en application n'aurait guère été possible sans ces outils, vu le goulot d'étranglement dû au volume énorme de connaissances nécessaires. Nos méthodes se basent autant que possible sur une analyse automatique des corpus source et cible, même si nous n'adoptons pas d'office les nouvelles techniques de pointe encore en développement.

Il apparaît clairement qu'en attendant la preuve de certaines technologies textuelles, il est déjà possible d'intégrer l'informatique et l'expertise humaine pour puiser le contenu de tels corpus. Même la vaste tâche d'élaboration d'une base de connaissance compréhensible pour tout un domaine technique n'échappe à une telle collaboration.

Cette dernière exige au moins une panoplie d'outils qui assurent le remaniement des fichiers de textes, l'analyse grammaticale et statistique des mots et phrases, l'affichage de termes dans leurs contextes unilingues et bilingues, et l'interaction efficace avec les experts humains. Notre contribution montre que de telles entreprises sont faisables.

9. Remerciements

Je tiens à remercier d'abord tous mes collègues du projet général KANT: James Altucher, Kathy Baker, Nicholas Brownlow, Alex Franz, Sue Holm, Kathy Iannamico, Pam Jordan, Todd Kauffmann, Marion Kee, Daniela Lonsdale, Mark Miller, Teruko Mitamura, Will Walker, nos associés chez Caterpillar et Carnegie Group, et notre directeur Jaime Carbonell. Ont participé notamment dans les tâches liées au vocabulaire français: Youcef Dahmane, Bertrand Damiba, Henry Debusmann, Kristina Keenan, Sarah Law, Jeanne Mier, et Hanming Ong. Surtout j'aimerais souligner la collaboration et les contributions indispensables de John Leavitt et Eric Nyberg du projet KANT, et particulièrement de Claude Doré de Traductions Taurus.

References

[Bauer, 1978] Bauer, L. (1978). *The grammar of nominal compounding with special reference to Danish, English, and French.* Odense University Press.

[Bédard, 1986] Bédard, C. (1986). *La traduction technique: principes et pratique.* Linguatech.

[Church, 1993] Church, K. W. (1993). Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics.*

[Debili and Sammouda, 1992] Debili, F. and Sammouda, E. (1992). Appariement des phrases de textes bilingues. In *Actes de COLING 1992.*

[Gale and Church, 1991] Gale, W. and Church, K. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.*

[Galinski, 1988] Galinski, C. (1988). Advanced terminology banks supporting Knowledge-Based MT. In Maxwell, D., Schubert, K., and Witkam, A. P. M., editors, *New Directions in Machine Translation.* Foris Publishers.

[Goodman and Nirenburg, 1991] Goodman, K. and Nirenburg, S., editors (1991). *A Case Study in Knowledge-Based Machine Translation.* Morgan Kaufmann.

[Levi, 1978] Levi, J. N. (1978). *The Syntax and Semantics of Complex Nominals.* Academic Press.

[Nyberg and Mitamura, 1992] Nyberg, E. and Mitamura, T. (1992). The KANT system: Fast, accurate, high-quality translation in practical domains. In *Actes de COLING 1992.*

[Savoy, 1991] Savoy, J. (1991). Stemming of French words. Département d'informatique et de recherche opérationnelle 793, Université de Montréal.

[Simard et al., 1992] Simard, M., Foster, G. F., and Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In *Actes de TMI-92.*

[Warren, 1978] Warren, B. (1978). *Semantic Patterns of Noun-Noun Compounds.* Acta Universitatis Gothoburgensis.

Abstract

This paper discusses our approach to the task of establishing a bilingual vocabulary for subsequent use in an automatic MT system. We outline methods for identifying, constraining, and aligning English source and French target vocabularies from pre-existing translation archives: source text analysis, single-word and phrasal context indices, identification of nominal compounds and variants, integration of client terminology resources, automatic alignment of source and target texts, extraction of translated source terms, vocabulary revision, and conversion for MT lexicon purposes. We also describe the tools designed and developed to help us in this effort: indexers, monolingual and bilingual context browsers, parsers, programs for describing nominal compound constituency, and a translation editor tool.