Meta
Journal des traducteurs
Translators' Journal

META

La traduction automatique au service de l'utilisateur monolingue

Laurence Jacqmin

Volume 37, Number 4, décembre 1992

Études et recherches en traductique / Studies and Researches in Machine Translation

URI: https://id.erudit.org/iderudit/003237ar
DOI: https://doi.org/10.7202/003237ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Jacqmin, L. (1992). La traduction automatique au service de l'utilisateur monolingue. *Meta*, *37*(4), 610–623. https://doi.org/10.7202/003237ar

Article abstract

This paper presents the Babel-Research Project on Machine Translation (MT) which started in January 1988. We recently completed a prototype which translates commercial letters from French into English. The MT system is intended for a monolingual secretary. This option imposed several constraints on the system design. Since the user is not able to correct the target text, the system must provide a satisfactory translation quality. One of our research lines is to examine whether a normative pre-edition and an interactive MT system offer a valid alternative to post-edition systems and what syntactic and semantic scope this kind of systems could comparatively propose. The paper presents our theoretical options: transfer architecture, computational modularity, unification paradigm. Then, it describes our prototype, Babel-2, and focuses on its main original aspects. First, we set up a double-step parsing strategy which combines the advantages of the Top-Down and Bottom-Up methods while avoiding their drawbacks. In short, the first movement deals with constituency, the second one deals with dependency. Secondly, the prototype recognizes source language idioms and translates them adequately. Third, it handles a number of non-trivial translational problems like aspectual nuances, saxon genitive construction, word order changes, as well as problems involving structural changes such as argument deletion, argument addition, dependency inversion... The paper ends with a glance at future research.

LA TRADUCTION AUTOMATIQUE AU SERVICE DE L'UTILISATEUR MONOLINGUE

LAURENCE JACQMIN
Université Libre de Bruxelles, Bruxelles, Belgique

Résumé

L'article présente le projet de recherche en traduction automatique, Babel-Research, qui a débuté en janvier 1988 sous l'égide de l'Université Libre de Bruxelles. Il a récemment abouti à la réalisation d'un prototype de traduction de correspondance commerciale français-anglais.

Le système est conçu comme un outil de travail pour un(e) secrétaire monolingue, ce qui impose certaines contraintes à l'architecture du système. Puisque l'utilisateur n'est pas à même de corriger le texte cible, le système doit garantir une qualité de traduction satisfaisante. Notre principale voie de recherche consiste à évaluer d'une part si un système basé sur la pré-édition normative du texte source et l'interaction avec son rédacteur en cours de traduction offre une solution valable aux systèmes à postédition plus courants ; et à évaluer d'autre part quelle couverture syntaxique et sémantique ce type de système peut comparativement proposer.

L'article présente d'abord les options théoriques du projet : architecture de transfert, modularité computationnelle, paradigme d'unification. Ensuite, il décrit le prototype, Babel-2, en mettant l'accent sur ses caractéristiques originales. Premièrement, nous avons mis au point une stratégie d'analyse en «pas de deux», qui combine les avantages des stratégies descendante et ascendante tout en contournant leurs inconvénients. En bref, un premier mouvement établit les relations de constituance, un second les liens de dépendance. Deuxièmement, Babel-2 reconnaît et traduit de manière adéquate divers types d'idiomes français. Troisièmement, il résout une série de problèmes de traduction complexes tels les nuances aspectuelles, le génitif saxon, les changements d'ordre des mots. Il traite également les constructions qui présentent des transformations structurelles telles que suppression d'arguments, adjonction d'arguments, inversion des relations de dépendance...

Un regard prospectif vers les recherches futures clôture la présentation.

Abstract

This paper presents the Babel-Research Project on Machine Translation (MT) which started in January 1988. We recently completed a prototype which translates commercial letters from French into English. The MT system is intended for a monolingual secretary. This option imposed several constraints on the system design. Since the user is not able to correct the target text, the system must provide a satisfactory translation quality. One of our research lines is to examine whether a normative pre-edition and an interactive MT system offer a valid alternative to post-edition systems and what syntactic and semantic scope this kind of systems could comparatively propose.

The paper presents our theoretical options: transfer architecture, computational modularity, unification paradigm. Then, it describes our prototype, Babel-2, and focuses on its main original aspects. First, we set up a double-step parsing strategy which combines the advantages of the Top-Down and Bottom-Up methods while avoiding their drawbacks. In short, the first movement deals with constituency, the second one deals with dependency. Secondly, the prototype recognizes source language idioms and translates them adequately. Third, it handles a number of non-trivial translational problems like aspectual nuances, saxon

genitive construction, word order changes, as well as problems involving structural changes such as argument deletion, argument addition, dependency inversion... The paper ends with a glance at future research.

INTRODUCTION

Le projet de recherche en traduction automatique, Babel-Research[1], a débuté en janvier 1988 sous l'égide de l'Université Libre de Bruxelles. Il a récemment abouti à la réalisation d'un prototype de traduction de correspondance commerciale dans le sens français-anglais.

Le système est conçu comme un outil de travail pour un(e) secrétaire monolingue, ce qui impose certaines contraintes sur l'architecture du système. Puisque l'utilisateur n'est pas à même de corriger le texte cible, le système doit garantir une qualité de traduction satisfaisante. Notre principale voie de recherche consiste à évaluer d'une part si un système basé sur la pré-édition normative du texte source et l'interaction avec son rédacteur en cours de traduction offre une solution valable aux systèmes à post-édition plus courants ; et à évaluer d'autre part quelle couverture syntaxique et sémantique ce type de système peut comparativement proposer.

L'article présente d'abord les options théoriques du projet : architecture de transfert, modularité computationnelle, paradigme d'unification. Ensuite, il décrit le prototype, Babel-2, en mettant l'accent sur ses caractéristiques originales. Premièrement, nous avons mis au point une stratégie d'analyse en «pas de deux», qui combine les avantages des stratégies descendante et ascendante tout en contournant leurs inconvénients. En bref, un premier mouvement établit les relations de constituance, un second les liens de dépendance. Deuxièmement, Babel-2 reconnaît et traduit de manière adéquate divers types d'idiomes français. Troisièmement, il résout une série de problèmes de traduction complexes tels les nuances aspectuelles, le génitif saxon, les changements d'ordre des mots. Il traite également les constructions qui présentent des transformations structurelles telles que suppression d'arguments, adjonction d'arguments, inversion des relations de dépendance...

Un regard prospectif vers les recherches futures clôture la présentation.

1. CARACTÉRISTIQUES EXTÉRIEURES DU PROJET

Parmi les systèmes de traduction automatique (TA) arrivés à maturité (état qui se concrétise par leur commercialisation), deux grandes familles se distinguent : d'une part, on trouve des systèmes portant sur un sous-langage très restreint et bien défini (par exemple : la météorologie, voir Kittredge et Lehrberger 1982). Ces systèmes fonctionnent de manière totalement automatique et fournissent une traduction de qualité satisfaisante. Mais la plupart des systèmes actuels, plus justement baptisés logiciels d'aide à la traduction (ou Traduction Assistée par Ordinateur, en abrégé TAO), s'appliquent à des domaines plus larges (documentation scientifique et technique) et se limitent à une traduction brouillon destinée à être relue et corrigée par un traducteur humain.

Notre système trace une troisième voie médiane. Entre la TA pour sous-langage et la TAO pour traducteurs, sa conception se décalque sur un demandeur de traduction jusqu'ici négligé : l'utilisateur monolingue, dont le besoin de traduction est régulier mais limité. Avec le domaine d'application, la correspondance commerciale, ce profil motive l'architecture de notre système.

1.1. DOMAINE D'APPLICATION : LE COURRIER COMMERCIAL

C'est principalement la recherche d'un produit destiné à enrichir une station de travail bureautique qui justifia le choix de la correspondance commerciale comme domaine d'application. À cette motivation, il faut joindre l'internationalisation croissante des

échanges commerciaux ainsi que l'affirmation largement diffusée par les manuels de correspondance que le courrier commercial constitue un sous-langage.

Après le tristement célèbre rapport ALPAC (1966) et l'éclipse qu'il a provoquée durant les années suivantes, la recherche en TA doit en partie son renouveau à la notion de sous-langage : l'histoire a montré qu'en réduisant ses ambitions linguistiques et traductiques aux dimensions d'un sous-langage, on pouvait espérer réaliser des systèmes de traduction totalement automatiques et de bonne qualité (ex. TAUM-METEO voir Nirenburg 1987 : 30-33 et Kittredge 1982).

Un sous-langage présente un ensemble de restrictions linguistiques qui rendent viables des applications informatiques :

1) Simplicité de la grammaire d'un sous-langage par rapport à celle de la langue prise dans sa totalité. Les constructions syntaxiques sont moins variées et moins complexes.
2) Réduction de la couverture lexicale (dans le cas de la correspondance commerciale, nous avons évalué la taille du lexique de base à 2 000 mots). Cette limitation facilite la maintenance et garantit la cohérence des dictionnaires du système. Elle promet également une sémantique plus circonscrite et rend possible l'énumération exhaustive des «restrictions de sélections», c'est-à-dire le type et le nombre des arguments verbaux (et même parfois nominaux et adjectivaux).
3) Existence d'une communauté de spécialistes faisant autorité, capable d'établir une typologie cohérente et exhaustive des concepts, des objets et de leurs interactions dans leur domaine de spécialité. L'univocité des définitions réduit l'ambiguïté lexico-sémantique.
4) Le sous-langage se caractérise par sa stabilité interlangue. Pour un même domaine, les différentes langues expriment les mêmes idées grâce à des moyens linguistiques similaires. Cette proximité facilite l'établissement des équivalences traductiques dans le système de TA.

Pour déterminer si le langage utilisé dans un champ de connaissances particulier correspond à un sous-langage, on examine, dans un premier temps, ce domaine d'un point de vue pragmatique :

1) S'agit-il d'un domaine restreint d'activité, de la description d'une partie limitée de la réalité ?
2) Existe-t-il une communauté de spécialistes faisant autorité dans le domaine, à même de produire une représentation cohérente et exhaustive des données ?
3) Les objets et les relations ont-ils reçu une définition univoque et universelle ?
4) Trouve-t-on des similitudes linguistiques dans des langues différentes, pour le domaine considéré ?
5) Enfin, dans l'optique d'une implémentation, l'ordinateur constitue-t-il un média privilégié pour ce champ d'activité ?

Dans un deuxième temps, l'étude linguistique d'un corpus représentatif de textes confirme (ou infirme) l'approche pragmatique. Cette étude s'attache à décrire en détail les restrictions lexicales, syntaxiques et sémantiques, les expressions figées du sous-langage considéré.

Nous avons appliqué cette grille d'étude (voir Jacqmin 1989) à la correspondance commerciale qui se présentait, à première vue, comme un bon candidat. En effet, le courrier commercial concerne un nombre restreint de situations qui partagent les mêmes objectifs, la même progression argumentative, les mêmes protagonistes : demande d'informations, livraison de marchandises, facture, réclamation... De plus l'internationalisation des transactions commerciales accroît le «métissage» linguistique. Ceci se révèle

particulièrement vrai pour les langues européennes telles que le français, l'anglais, l'allemand et le néerlandais. Enfin, l'ordinateur constitue un outil essentiel pour le monde commercial.

Mais un regard plus attentif nous mena à la conclusion inverse. Nous n'avons pas pu identifier une communauté de spécialistes faisant autorité dans le domaine, ni découvrir une typologie conceptuelle reconnue. De plus, le niveau d'expertise hétérogène des rédacteurs de courrier commercial (du directeur commercial à la secrétaire dactylo) se marque sur le style, la complexité grammaticale et la précision terminologique des lettres.

Deuxièmement, nous avons passé au crible un corpus représentatif de lettres commerciales, certaines réelles, d'autres émanant de manuels de correspondance. Contrairement à ce que la plupart des manuels affirment, les lettres présentent une complexité syntaxique (interrogatives, impératives, subordinations en cascade, discours indirect, tournures impersonnelles...) et une richesse sémantique non négligeables.

Parce que le commerce touche à tous les aspects de la vie, son lexique présente un horizon fuyant, difficile à délimiter a priori. Il en est de même pour l'ambiguïté lexicale qui s'y apparente. Par exemple, le mot *facture* : à première vue, on supposera qu'il réfère au document comptable. Mais dans la correspondance d'un facteur d'instruments, il peut aussi caractériser la facture d'un instrument[2].

En l'absence d'un sous-langage commercial suffisamment restreint, nous avons sélectionné au sein de notre corpus un noyau à la fois représentatif du domaine et linguistiquement simplifié, ce qui revient à substituer à une approche descriptive du sous-langage une approche prescriptive. Cette dernière s'inscrit d'ailleurs dans une tendance actuelle (voir Blanchon 1990) à la normalisation de la communication institutionnelle régie par des règles de rédaction, des modèles linguistiques susceptibles d'accroître l'efficacité de la communication et de faciliter le traitement des informations[3].

1.2. SYSTÈME INTERACTIF AVEC PRÉ-ÉDITION

Dans son survol de l'histoire de la TA, Boitet (Boitet 1989 ; voir aussi Somers 1991) distingue trois types de systèmes suivant le profil de l'utilisateur : les systèmes destinés aux «sentinelles» rendent accessibles le contenu principal du texte plus qu'ils ne le traduisent (SYSTRAN, voir King 1987), ensuite viennent les systèmes conçus pour les réviseurs (METAL, voir Nirenburg 1987 ; EUROTRA, voir King 1987), les systèmes à postédition, enfin les systèmes d'aide aux traducteurs professionnels qui proposent une série d'outils interactifs. Ce troisième type majoritairement représenté sur le marché actuel (ALPS, WEIDNER... *in* King 1987), offre surtout des traitements de textes perfectionnés auxquels sont associés d'énormes dictionnaires bilingues, des correcteurs orthographiques, des dictionnaires de synonymes... Outre la garantie d'une plus grande cohérence lexicale (le même mot recevra toujours la même traduction), leur intérêt se réduit à leur vitesse de traduction, qui peut diminuer de moitié le temps de travail global (TA + correction humaine) pour plusieurs milliers de pages.

Face à ces utilisateurs, tous bilingues, émerge un autre utilisateur potentiel de TA : l'utilisateur monolingue, dont le besoin de traduction est régulier mais trop limité pour faire appel aux services d'un traducteur professionnel.

Cet utilisateur a suscité l'intérêt de plusieurs équipes de recherche (ENtran développé chez UMIST, voir Nirenburg (1987), Somers (1991) ; Rosetta, voir Sanders (1988)) parmi lesquelles BABEL-Research.

Notre prototype, Babel-2, cible un utilisateur étranger à la pratique de la traduction et sans connaissance linguistique particulière : un(e) secrétaire écrit une lettre dans sa langue maternelle et désire en obtenir la traduction dans une langue qu'il ne connaît pas, ou dans laquelle il rédige avec peine.

Cette option pèse sur l'architecture du système : un utilisateur qui ignore la langue cible (LC) ne peut réviser le texte après traduction. Le système automatique doit garantir une qualité de traduction satisfaisante.

Notre système s'applique à un domaine trop large pour exclure toute intervention humaine du processus, mais la postédition ne convient bien sûr pas à notre utilisateur monolingue. Une de nos voies de recherche consiste primo : à examiner si une pré-édition normative, jointe à une interaction en cours de traduction, en langue source (LS) exclusivement, constitue une solution valable aux systèmes avec postédition, secundo : à évaluer la qualité du résultat et la couverture linguistique de ce type de systèmes.

Comme nous avons déjà évoqué l'aspect normatif du système, nous nous concentrerons maintenant sur l'interaction.

L'interaction vient en renfort pour résoudre le principal problème qui se pose à la TA : l'ambiguïté du langage naturel. Celle-ci peut prendre diverses formes : morphologique, syntaxique, sémantique, structurale ou lexicale. Pour la résoudre, l'interlocuteur humain fait intervenir de multiples sources d'informations linguistiques et extra-linguistiques (connaissance du monde, du locuteur, contexte d'apparition de la phrase). À ce jour, les interactions entre les différents types d'informations, leur fonction désambiguïsante n'ont pas reçu de description complète et satisfaisante ni d'un point de vue théorique ni d'un point de vue pratique (c'est-à-dire propre à déboucher sur une application). Quoi qu'il en soit, l'élaboration d'un tel système expert de désambiguïsation dépassait de loin nos ambitions de recherche. Nous avons préféré faire appel, comme d'autres systèmes d'ailleurs, à l'expertise humaine pour résoudre les ambiguïtés résiduelles en fin d'analyse (*cf.* Rosetta[4]) et durant le transfert[5].

2. OPTIONS THÉORIQUES

2.1 ARCHITECTURE DE TRANSFERT

Dans un premier temps, le projet s'est concentré sur le français et l'anglais, mais son architecture reflète l'intention d'étendre la portée du prototype à d'autres langues telles que le néerlandais et l'allemand. Dans ce contexte, une architecture basée sur le transfert s'avère plus appropriée. Un module de transfert spécifique à une paire de langues traite tous les phénomènes traductiques et fait le lien entre l'analyse en langue source (LS) et la génération en langue cible (LC). Ces deux modules, conçus de manière totalement monolingue, restent valides et disponibles en cas d'extension du système à d'autres langues.

On oppose habituellement cette architecture aux systèmes à interlangue qui produisent une représentation intermédiaire au contenu de type sémantico-logique, entre l'analyse et la génération (exemples : KBMT, voir Nirenburg 1987 : 136-142).

Outre des contre-arguments d'ordre théorique (certains doutent de l'existence d'une interlangue universelle), cette architecture ne tire pas profit de la stabilité linguistique qui caractérise le courrier commercial (*cf.* «métissage linguistique» évoqué plus haut), à laquelle s'ajoute la proximité des langues prises en compte.

Dans notre cas, une architecture de transfert gagne en profondeur d'analyse : un bon nombre d'ambiguïtés (ex. portée des quantifieurs) se transfèrent sans résolution, sans la nécessité d'un passage par une représentation profonde de nature sémantique, logique ou même pragmatique.

Le choix d'une architecture de transfert impose une structure tripartite au dictionnaire : un dictionnaire monolingue pour chaque langue traitée et un dictionnaire bilingue, le dictionnaire de transfert, spécifique à une paire de langues, comme le module du même nom.

2.2. MODULARITÉ

Dans la plupart des systèmes de TA, la modularité computationnelle calque la stratification linguistique communément admise : morphologie, syntaxe, sémantique. Nous y voyons le résultat d'une confusion entre modularité informatique et stratification linguistique. Certes, l'étude linguistique se doit de faire cette distinction conceptuelle, mais cela n'impose pas aux systèmes automatisés de la reproduire comme autant de phases disjointes et successives.

Dissocier l'accès aux informations linguistiques d'après leur nature nuit à l'efficacité du système. La production d'ambiguïtés s'élève considérablement et chaque phase n'est plus destinée qu'à filtrer les nombreux résultats non valides de la précédente.

En réalité, de nombreux phénomènes linguistiques ne peuvent être décrits que par le recours simultané à des informations de tout ordre. Par exemple, la référence à des dates apparaît fréquemment dans la correspondance commerciale, sous la forme de syntagmes nominaux postverbaux (voir [1]).

[1] Nous vous enverrons le 4 février prochain les 10 écrans commandés.

Des critères purement syntaxiques (catégorie, position relative dans la phrase...) font du syntagme *le 4 février prochain* l'objet direct du verbe. Par conséquent, pour toutes les phrases contenant une date à cet endroit, une analyse syntaxique produirait deux solutions, dont l'une serait systématiquement filtrée lors de l'analyse sémantique[6]. Le même problème surgit pour l'attachement des adjectifs postposés en français qui ne dépendent pas nécessairement, comme l'illustre [2], du nom qu'ils suivent directement.

[2] condition de livraison ci-jointe
où *ci-jointe* se rapporte à *condition*

Face à la conception en parallèle[7] se présente une approche plus abordable : l'accès simultané mais sélectif à tous les types d'informations.

Nous inspirant de formalismes récents tels que HPSG, *Head-driven Phrase Structure Grammar* (Pollard 1987), nous considérons, que dans l'usage comme dans l'interprétation du langage, divers types d'informations linguistiques et extra-linguistiques (*cf.* plus haut ambiguïtés) interviennent et interagissent. L'analyse d'une phrase revient en quelque sorte à déterminer la nature précise de ces interactions. La description déclarative et uniforme (par structures de traits) des faits et des contraintes linguistiques les rend accessibles en permanence. Grâce à un accès «à la carte» aux informations linguistiques utiles, chaque phase traite exhaustivement son unité de traitement, qui va du caractère à la structure prédicative de la phrase, en passant par le mot simple puis composé, et le syntagme (éventuellement composé lui aussi).

2.3. FORMALISME

Le développement du prototype a bénéficié d'un environnement d'unification de style PATR-II avec unification de structures, disjonction et négation atomiques (voir Shieber 1986). De plus, le prototype incarne certaines options du paradigme d'unification :

- Le module d'analyse produit une structure prédicative de surface.
- Les informations linguistiques sont représentées dans des graphes acycliques orientés ou structurés de traits, c'est-à-dire une structure récursive de paires attribut-valeur dont la valeur est soit un atome soit une structure de traits elle-même. Ce mode de représentation des données, qui décrit un objet par l'énumération de ses caractéristiques distinctives, rejoint des techniques d'intelligence artificielle telles que les scénarios (*scripts*) et les modèles (*frames*).

- L'unification constitue l'opération clé de manipulation des informations.
- Le caractère déclaratif du système (*vs* procédural) assure sa modularité.

À ces caractéristiques s'associe un mode de programmation logique (PROLOG).

3. LE PROTOTYPE

Nous l'avons déjà évoqué, l'absence de réel sous-langage commercial nous a contraints à imposer un certain nombre de restrictions sur la complexité linguistique admise par le système de TA. L'étude du corpus de lettres commerciales a permis de dégager un ensemble de constructions représentatif du courrier courant. Babel-2 traduit actuellement des lettres comme [3] :

[3] Monsieur,
Nous accusons réception de votre lettre du 20 novembre relative à notre commande numéro 1234 de circuits intégrés. Par cette lettre, vous nous informez de la livraison imminente du matériel de remplacement.
Dans cette lettre, vous réclamez un premier paiement rapide. Le 21 novembre dernier, nous avons viré sur le compte de votre firme, numéro 12-34567-02, la somme de 12 000 francs belges. Etc.

Babel-2 traite les phrases déclaratives affirmatives contenant un verbe conjugué. De plus, la subordination nominale est prise en compte dans toute sa complexité. Comme le montrent les exemples [4,5], la dépendance nominale implique des rattachements bien plus élaborés que de simples mouvements de droite à gauche.

[4] Nous acceptons le report de votre paiement au mois prochain.
Au mois prochain se rapporte à *report* et non à *paiement*.
[5] Nous vous enverrons le 5 décembre prochain nos conditions de livraison habituelles.

habituelles caractérise *conditions* et non *livraison*. Dans cet exemple précis, la vérification de l'accord morphologique entre nom et adjectif suffirait, mais souvent seules les contraintes sémantiques sont discriminantes pour l'attachement de l'adjectif postposé.

Babel-2 se distingue surtout par son traitement des composés et des locutions, et par la résolution d'une série de problèmes traductiques complexes.

3.1. MOTS COMPOSÉS ET LOCUTIONS

Nous distinguons trois types de composés d'après leur comportement morpho-syntaxique et leur degré de variabilité.

3.1.1. COMPOSÉS MORPHOLOGIQUES

[6] circuit intégré,
[7] au sujet de...

La variabilité de [6,7] se limite aux variations flexionnelles admises par la catégorie du composé ; un composé nominal varie en nombre et en genre, un composé verbal en temps, en mode, etc. Les éléments constitutifs du composé apparaissent contigus dans la phrase.

La phase de reconnaissance morphologique qui débute l'analyse regroupe les éléments du composé en une unité lexicale et lui associe sa description spécifique (*au sujet de* —> préposition). Pour ce faire, elle consulte une table des composés dans le dictionnaire, qui contient les contraintes de reconnaissance et les descriptions des composés.

3.1.2. COMPOSÉS SYNTAGMATIQUES

[8] sous huitaine,
[9] en bon / mauvais état
[10] accuser [bonne] réception
[11] en sa / votre / leur faveur
[12] soumettre à l'approbation dans la phrase : nous soumettrons ce projet de construction à l'approbation de notre Conseil d'administration.

Ces locutions portent sur un ou plusieurs syntagmes et leur variabilité est plus riche : elles peuvent présenter des éléments optionnels ([10] ; certains mots peuvent varier en degré ([9]), en personne ([11]). L'adjacence des composants n'est pas requise ([12]).

Entre la phase de regroupement des mots en syntagmes et celle-ci se fixent leurs relations de dépendance, une phase spécifique traite les composés syntagmatiques. Ici encore, le prototype tire tous les avantages de la lexicalisation. Le mot le plus discriminant d'un composé déclenche la consultation de la table des composés syntagmatiques qui décrit les contraintes de reconnaissance du composé (mots requis et leurs caractéristiques morpho-syntaxiques). Cette phase regroupe les syntagmes du composé éventuellement disséminés dans la phrase et réarrange la séquence des mots en conséquence. À la nouvelle unité lexicale complexe sont associées, au même titre qu'à toute unité lexicale, ses descriptions syntaxico-sémantiques, indispensables au succès de la phase de structuration qui suit.

3.1.3. COMPOSÉS DE TRANSFERT

[13] communication téléphonique —> *phone call*,
[14] au prix de 500 FB la pièce —> *at the unit price of FB 500* composé : prix — pièce

En l'absence de traduction mot à mot, le transfert regroupe les mots en langue source de manière à pouvoir leur associer un équivalent traductique global. Pour les cas comme [13,14], le dictionnaire de transfert contraint l'environnement de la tête syntagmatique. Une entrée lexicale établit l'équivalence traductique entre *communication* et *phone call* lorsque *communication* est modifié dans le texte source par *téléphonique*. Une deuxième entrée lexicale traite la traduction de *communication* sans contrainte lexicale sur l'environnement : *communication* (français) —> (anglais) *communication*.

Le traitement des composés de transfert suit le même schéma que celui des composés d'analyse. Dans le dictionnaire de transfert, la tête en LS équivalant à celle du syntagme idiomatique en LC active la recherche des autres éléments du composé. Si les contraintes de reconnaissance sont satisfaites, l'idiome en LC constitue l'équivalent traductique dans la structure résultat du transfert, qui peut éventuellement avoir subi des changements structurels liés à l'apparition de l'idiome, comme l'illustre [15].

[15] en bois —> *wooden*

Le syntagme prépositionnel (SP) français devient un adjectif en anglais. Durant la phase de transfert, la tête *bois* du SP déclenche un processus qui vérifie la présence exclusive de *en* dans le SP (*cf. en bois d'olivier —> made of olive wood*) et effectue les modifications structurelles : l'adjectif s'intègre au syntagme nominal (SN) dont le SP dépendait en français.

3.2. PROBLÈMES TRADUCTIQUES

Pour assurer une bonne qualité de traduction, Babel-2 a l'ambition de dépasser le mot à mot et traite un certain nombre de divergences traductiques qui suscitent des changements structurels.

3.2.1. LA MODIFICATION DE L'ORDRE DES MOTS

Le module de génération se base non pas sur la séquence de mots en LS (français) mais sur un ensemble de règles de préférence propres à la langue cible (anglais) mettant en jeu des critères d'ordre sémantique ou thématique (les rôles occupés par les syntagmes) vis-à-vis de leur tête.

[16] Notre client vous a acheté 50 actions BBL.
Our customer has bought 50 BBL shares from you.

Dans [16], le bénéficiaire pré-verbal pronominalisé *vous* devient un syntagme prépositionnel postverbal *from you*; *BBL*, complément nominal spécifiant le nom ou l'identité, est postnominal en français et pré-nominal en anglais.

On trouve aussi des modifications internes aux syntagmes. Par exemple, la plupart des adjectifs postposés en français sont pré-posés en anglais.

3.2.2. LA MODIFICATION CATÉGORIELLE

Elle se présente avec certaines nominalisations françaises dans le cas d'enchâssements en cascade.

[17] Nous acceptons votre proposition de report du paiement au mois prochain.
We accept your proposal to defer the payment until next month.

Dans le contexte syntaxique particulier illustré par [17], l'équivalent verbal *to defer* traduit le nom français *report*. Le module de transfert contient une règle qui vérifie le nombre d'enchâssements et modifie, le cas échéant, la catégorie du nom pris en sandwich. La sélection de l'équivalent traductique suit.

Comme le montre [18], la catégorie subit aussi des changements en cas d'absence d'équivalent direct.

[18] Nous fournirons la documentation à votre firme bruxelloise.
We will supply the information to your firm in Brussels.

Une locution spatiale *in Brussels* traduit l'adjectif français *bruxellois*. Cette transformation a lieu durant le processus de transfert. Dans le dictionnaire de transfert, l'unité lexicale *bruxellois* active une règle de transfert qui ajoute un complément de lieu à la structure de dépendance (et supprime l'adjectif).

3.2.3. LE GÉNITIF SAXON

[19] Nous déposerons le colis de vos clients à la banque dès réception de nouvelles directives.
On receipt of new instructions we will leave your customers' parcel at the bank.

Le générateur ordonne les syntagmes sur la base de leurs caractéristiques sémantiques. Lorsque le modifieur d'un syntagme nominal appartient à une certaine classe sémantique (intervenants humains et assimilables), le générateur introduit la construction à génitif saxon, qui provoque plusieurs changements structurels : i) une règle de préférence anglaise spécifique à la construction génitive détermine le nouvel ordre des mots, ii) le déterminant de la tête disparaît, iii) la marque du génitif ('s ou ' ou encore s') est placée derrière la tête du complément antéposé.

3.2.4. LE TEMPS ET L'ASPECT VERBAUX

Entre français et anglais, le temps ne se traduit pas toujours de manière directe.

[20] La firme vous réclame la livraison des écrans depuis Noël.
As from Christmas the firm has claimed the delivery of the screens from you.

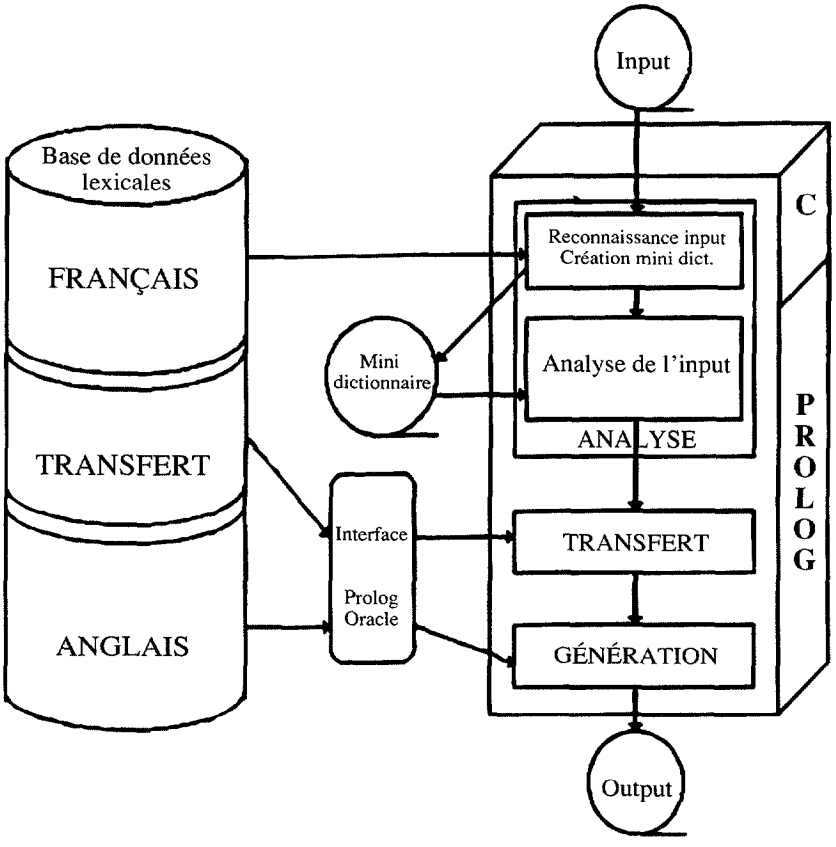
Dans [20], le présent *réclame* se traduit par un *present perfect* en anglais *has claimed*. Le module de transfert détecte l'aspect duratif grâce à la présence du complément de temps *depuis Noël*. Il sélectionne ensuite le temps correspondant en anglais.

3.2.5. LES NOMS COMPTABLES / NON COMPTABLES

[21] les trois meubles
the three pieces of furniture

Durant le transfert, l'équivalence est établie entre *meuble* et *furniture*. Une règle de transfert, activée par la différence comptable/non comptable entre les deux équivalents, ajoute un syntagme qui agit comme quantifieur tel que *piece of*, dans [21][8] et réorganise le syntagme en conséquence.

3.3. FONCTIONNEMENT DU TRADUCTEUR



Le traducteur se compose principalement d'une base de données relationnelle (gérée sous ORACLE[9]), tripartite, contenant l'ensemble des informations lexicales, et d'une séquence de processus d'analyse, de transfert et de génération. Au début de l'analyse, un mini-dictionnaire concentre les descriptions lexicales de tous les mots de la lettre

à traduire en éliminant les redondances éventuelles (apparitions multiples d'un mot). Ensuite, pour les accès ponctuels au lexique, le traducteur recourt à une interface PROLOG-ORACLE développée sur mesure dans le cadre du projet BABEL-R.

3.3.1. LE LEXIQUE

La base de données lexicales décrit actuellement 1 025 mots, regroupés en 250 unités lexicales. Celles-ci servent de pivot entre les flexions d'un mot et sa ou ses description(s) syntaxico-sémantique(s). Une typologie de classes sémantiques permet de caractériser chaque mot. De plus, le lexique décrit ses environnements admis, c'est-à-dire ses arguments et/ou modifieurs.

La richesse descriptive du lexique agit comme une «main courante» durant toute l'analyse. Elle peut aussi dans certains cas orienter le choix d'équivalents traductiques

[22] demande —> *demand* ou *request*, selon les cas.

3.3.2. UNE ANALYSE «EN PAS DE DEUX»

Nous avons mis en œuvre une stratégie mixte qui combine les avantages des deux méthodes descendante (*Top-down*) et ascendante (*Bottom-up*) et évite leurs inconvénients.

La phase d'analyse se développe en deux mouvements. Le premier, descendant, élabore une structure syntaxique, fortement aplatie de la phrase. Il regroupe les mots en syntagmes et contraint la combinatoire de ceux-ci sans les organiser entre eux.

[23] Nous vous enverrons le 5 décembre prochain nos conditions de livraisons exceptionnelles.
We will send you our special delivery conditions on next 5 December.

La phrase [23] sera d'abord représentée comme une suite de syntagmes simples: Pronom Pronom SV SN SN SP AP. Cette phase prend aussi en charge la séquence de clitiques et l'application de contraintes morphologiques et configurationnelles permet d'identifier *nous* comme sujet, et *vous* comme l'oblique. Un processus spécifique aux dates analyse *le 5 décembre prochain*. Il combine des contraintes syntaxiques et sémantiques (dans le cas des dates, l'adjectif postposé est rattaché dès la première phase de l'analyse, ce qui n'est pas le cas pour les autres adjectifs [*cf. exceptionnelles*]).

Le deuxième mouvement, ascendant, construit une structure de dépendance de la phrase. La confrontation des valences verbales, nominales et adjectivales et des contraintes fonctionnelles, lexicales et sémantiques qu'elles contiennent permet d'élaborer la structure prédicative de la phrase, représentée sous forme de structure de traits (voir [24]).

[24] Pour faciliter la lecture de la struture de traits, nous ne faisons figurer que les informations pertinentes.

```
[tête      envoyer,
  arg1 :[      tête : 'nous',
               cas : agent],
  arg2 :[      tête : 'vous',
               cas : bénéficiaire],
  arg3 :[      tête : 'conditions'
               cas : patient,
               composants : [det : 'nos',
                             adj : 'habituelles']
               arg3 : [   tête : 'livraison',
                          cas : patient]]
```

Cette stratégie permet de construire dynamiquement une structure de dépendance plutôt que d'éliminer les nombreuses structures qui auraient été élaborées à partir de contraintes purement configurationnelles.

L'aplatissement de la structure, conjugué à l'utilisation d'une table de solutions partielles (*well-formed substrings table*), élimine les inconvénients de la méthode descendante évoqués plus haut (retour en arrière coûteux et multiples élaborations de la même analyse partielle). De plus, le premier mouvement prévient la création d'analyses partielles incohérentes.

3.3.3. TRANSFERT LEXICAL

Le lexique télécommande le transfert. En effet, le choix des équivalences se fait par consultation du dictionnaire de transfert qui spécifie pour chaque mot, dans un sens ou dans un environnement donné, son équivalent en langue-cible. Le dictionnaire pose éventuellement des contraintes sur le contexte d'apparition du mot en langue-source. À ces contraintes, il associe, selon les cas, des instructions qui activent durant la phase de transfert des procédures de transformations structurelles[10], comme la suppression d'arguments en [24], l'ajout d'arguments en [26], l'intégration d'un argument dans un syntagme en [27], la modification catégorielle (*cf.* exemple [17]), la dépendance inversée en [28].

[25] erreur de calcul —> *miscalculation*
[26] échantillonnage —> *range of samples*
[27] boîte en bois —> *wooden box*
[28] travail accru —> *increased amount of work*

3.3.4. LA GÉNÉRATION

Le module de génération reçoit en input une structure de dépendance thématique (le verbe et ses différents arguments) contenant les équivalents anglais des mots et locutions français.

Après l'avoir enrichie des informations lexicales anglaises (essentiellement morpho-syntaxiques), le module organise la séquence d'arguments sur la base des règles de préférence anglaise, sans référence à l'ordre initial en LS. Ces règles, basées sur des critères sémantiques (cas et classes sémantiques des têtes), dictent l'ordre relatif des constituants. Enfin, la catégorie des mots régit l'ordre intra-syntagmatique.

4. CONCLUSION

Placé sous l'égide de l'Université Libre de Bruxelles, BABEL-Research se présente comme le fruit d'une collaboration entre Université et Industrie. Ce contexte mixte a créé une tension stimulante au sein du projet. La composante académique y a apporté ses exigences théoriques. Refusant de sacrifier à l'adage «la fin justifie les moyens», elle a assuré au projet des assises valides: architecture de transfert, modularité, formalisme d'unification. D'autre part, le partenaire industriel donnait à la recherche un cadre réaliste, en termes de cible: un utilisateur monolingue, en termes d'environnement: une station de travail bureautique, en termes de domaine d'application: la correspondance commerciale. Il imposait la réalisation d'un prototype, qui reste, en linguistique informatique, le mode de validation le plus implacable.

Babel-2 doit encore subir quelques améliorations:

1) Fusion des deux mouvements d'analyse. Déjà en cours, elle offre des résultats prometteurs: la mémorisation des solutions partielles diminue le coût de traitement (temps et espace mémoire).
2) Lexique dynamique. Comme mentionné plus haut, nous avons l'intention de rapprocher l'implémentation du lexique de son organisation conceptuelle de manière à prendre en compte les phénomènes d'héritage (simple ou multiple) et les généralisations lexicales.
3) Couverture étendue. La grammaire traitera les relatives et les complétives, la coordination comme la négation.

Ces aménagements vont de pair avec une réflexion plus théorique menée actuellement par les chercheurs sur :

1) la sémantique lexicale, dédiée à la traduction automatique : représentation, organisation et formalisation des données lexico-sémantiques dans le lexique ;
2) les outils et les formalismes de description linguistique, et plus précisément les interactions qui s'y tissent entre les notions de constituance, de dominance et de précédence linéaire.

Notes

1. En abrégé, BABEL-R. Les lecteurs de Tintin se souviendront du cheik Bab-el-ehr (dont le nom désigne, en bruxellois, le bavard intempestif). L'allusion biblique à la tour du même nom acheva de convaincre les initiateurs du projet qui fut baptisé «Babel-R(esearch)».
2. Pour cette raison, nous avons prévu la possibilité pour tout système commercialisé à partir de notre prototype d'étendre le lexique au vocabulaire spécialisé de l'utilisateur.
3. Le succès grandissant de la norme SGML (*Standard Generalized Marking Language*) illustre à souhait ce courant.
4. Landsbergen (1988), Sanders (1988) et Appelo (1986) évoquent le recours à l'interaction en langue source avec le rédacteur pour désambiguïser les cas insolubles pour le système. À ma connaissance, cet aspect de Rosetta, souvent évoqué, n'a cependant jamais fait l'objet de publications détaillées et accessibles.
5. Cet aspect de la recherche en est toujours à l'état de projet, car il convient avant tout de répertorier ces ambiguïtés (parfois créées par le système), d'identifier leur nature, leur fréquence, mais aussi de mettre au point l'interface qui dialoguerait avec l'utilisateur.
6. Il serait imprudent de classer ce problème au rang des phénomènes d'envergure négligeable, car la complexité et la longueur des phrases de notre corpus amenaient une analyse purement syntaxique à produire jusqu'à 4 000 arborescences pour une phrase !
7. Une approche en parallèle tente de contourner le problème (une stratégie blackboard a été envisagée à BABEL-R), mais ces systèmes sont compliqués, difficiles à maîtriser et très gourmands. Jusqu'à présent ils ont posé plus de problèmes qu'ils n'en ont résolus.
8. Le choix du quantifieur est lexicalisé. L'entrée lexicale monolingue du mot non comptable précise le quantifieur requis, par exemple *a lump of* pour *sugar*.
9. Ce sont principalement des raisons de confort au chargement et à la maintenance informatique de la BD qui ont motivé le choix du SGBD relationnel ORACLE. Mais il a imposé une conception statique de la lexicographie, interdisant toute possibilité d'y intégrer des généralisations linguistiques ou des phénomènes d'héritage. Cet aspect du prototype est en discussion.
10. F. van Dooren (1992) a montré qu'un traitement adéquat des divergences traductiques requiert une étude linguistique globale. Sur la base d'une typologie des divergences, linguistiquement motivée, on peut déterminer les informations que manipule chaque type de divergence et le niveau linguistique du module de transfert le plus approprié pour leur résolution.

BIBLIOGRAPHIE

ALPAC (1966) : *Language and Machines : Computers in Translation and Linguistics. A Report by the Automatic Language Processing Advisory Committee (ALPAC)*, Division of Behavioral Sciences, National Academy of Sciences, National Research Council Publication 1416, Washington, NAS / NRC.

APPELO, L. et J. LANDSBERGEN (1986) : «The Machine Translation Project Rosetta», *Proceedings of the International Conference on State of Art in MT*, Saarbrücken, pp. 34-51.

BLANCHON, H (1990) : «LIDIA-1, un prototype de TAO personnelle pour rédacteur monolingue», *Specialized Conference on Natural Language Processing & its Applications*, (Avignon, 1990), pp. 51-60.

BOITET, C. (1989) : «Speech Synthesis and Dialogue Based Machine Translation», *ATR Symposium on Basic Research for Telephone Interpretation*, Kyoto.

GAZDAR, G. et C. MELLISH (1989) : *Natural Language Processing in Prolog. An Introduction to Computational Linguistics*, Wokingham, Addison-Wesley Publishing Company.

JACQMIN, L. (1989) : *La notion de sous-langage et son importance pour la traduction automatique*, Technical Report BABEL-Research.

JOHNSON, R. L. et P. WHITELOCK (1987) : «Machine Translation as an Expert Task», *Machine Translation. Theoretical and Methodological Issues*, Nirenburg S. (Ed.), Studies in Natural Language Processing, Cambridge, Cambridge University Press, pp. 136-144.

KING, M. (Ed.) (1987) : *Machine Translation Today*, EDITS, vol. 2, Edinburgh, Edinburgh University Press.

KITTREDGE, R. et J. LEHRBERGER (Eds.) (1982) : *Sublanguage. Studies of Language in Restricted Semantic Domains*, Berlin, Walter de Gruyter.

LANDSBERGEN, J. (1988) : «Dictionaries for Rosetta», *Proceedings of the International Symposium on Electronic Dictionary*, Tokyo.

NIRENBURG, S. (Ed.) (1987) : *Machine Translation. Theoretical and Methodological Issues*, Studies in Natural Language Processing, Cambridge, Cambridge University Press.

PEREIRA, F. C. N. et S. M. SHIEBER (1987) : *Prolog and Natural-Language Analysis*, CSLI, Lecture Notes, number 10, Stanford.

POLLARD, C. et I. A. SAG (1987) : *Information-Based Syntax and Semantics*, CSLI, Lecture Notes, number 13, Stanford.

SANDERS, M. J. (1988) : *The Rosetta Translation System*, Eindhoven, Philips Research Laboratories, M.S.15.103.

SHIEBER, S. M. (1986) : *An Introduction to Unification-Based Approaches to Grammar*, CSLI, Lecture Notes, number 4, Stanford.

SOMERS, H. (1991) : «Current Research in Machine Translation», to appear in *Machine Translation*.

VANDOOREN, F. (à paraître) : «Un exemple de problème syntaxique : les divergences de traduction entre l'anglais et le français», *Manuel de traduction automatique* (titre provisoire), P. Bouillon (Éd.).