Marie-Hélène Bouveret-Mounpain

DIRECTEUR ADJOINT DU CLA/BESANÇON
SECTION DE FRANÇAIS

Aventures au pays de la traduction automatique

La Compagnie « LOGOS COMPUTER SYSTEMS INC. » a été créée en 1969 dans le but de développer des systèmes de traduction automatique. Au début le financement principal provenait de l'armée américaine : le premier système était le système fonctionnant de l'anglais au vietnamien (1970-1972, réalisé au bénéfice de l'armée de l'air et qui a traduit 5 millions de mots anglais en vietnamien avant que les événements politiques ne mettent un terme prématuré à sa courte existence).

Puis on a assisté à la création des systèmes anglo-russe puis anglo-persan dont la réalisation a rencontré quelques obstacles de type financier.

En 1979, la compagnie commence à travailler sur une traduction de l'allemand vers l'anglais. Le projet est financé par la firme allemande Siemens. Il se développe rapidement et devient bientôt le premier projet commercialisé par LOGOS en 1982. Le système fonctionne dans les deux sens allemand-anglais et anglais-allemand. Depuis cette date ont été développés différents systèmes : anglais-français et anglais-espagnol. En projet : traduction de l'allemand vers le français, de l'anglais vers l'italien et vers le portugais.

Les principaux utilisateurs de LOGOS à ce jour sont : la CEE à Luxembourg, Nixdorf, Opel, Siemens, Mercedes, Pfaff, l'armée américaine, en Allemagne, I.B.M. aux Etats-Unis, Burroughs aux Etats-Unis et en France.

Le système LOGOS est accessible sur des machines WANG ou IBM. On envisage d'étendre la gamme des utilisateurs potentiels dans un proche avenir.

L'équipe ou plutôt les équipes

L'élaboration du système qui va de l'anglais vers le français, comme de tous les autres systèmes d'ailleurs, était assumée par quatre équipes :

— Les linguistes : analysent le fonctionnement de la langue source et celui de la langue cible, les caractéristiques de chaque élément linguisti-

que, ajoutent continuellement de nouveaux raffinements dans les règles de fonctionnement au fur et à mesure qu'ils rencontrent des problèmes dans les tests continuels qu'ils effectuent (introduction d'un corpus de plus en plus étendu dans la machine et traduction par le système à tous les stades de son élaboration). Les linguistes sont également chargés d'expliquer leurs besoins aux programmeurs, d'en obtenir tout le système de codage (voir plus loin).

— Les programmeurs : tous anglophones et seulement anglophones pour la plupart, tous des petits génies de l'informatique qui exécutaient les ordres des linguistes mais qui parvenaient aussi à leur suggérer des possibilités diverses de codification permettant d'accélérer le processus de traduction. Le seul problème tenait à la mobilité de cette catégorie de personnel qui était vite débauchée par d'autres branches de l'industrie informatique qui leur faisait des ponts d'or dès qu'ils devenaient performants. La recherche dans ce secteur comme dans beaucoup d'autres aux Etats-Unis est très mal payée et la sécurité de l'emploi y est pratiquement inexistante.

— Le dictionnaire : cette équipe était chargée d'entrer un très grand nombre de mots (en 8 mois elle en a entré 28 000) avec une seule traduction par mot (la plus courante, bien entendu). C'est cette traduction « générique » qui est prise en compte à l'étape 2 de la traduction (voir plus loin). Cette équipe codifie toutes les caractéristiques de chaque mot introduit et de son équivalent dans la langue cible. Par exemple, pour un verbe, elle dira s'il est transitif, intransitif, pronominal réfléchi ou réciproque, s'il prend être ou avoir comme auxiliaire, etc. C'est encore cette équipe qui réalisera les dictionnaires de langue de spécialité qui seront proposés « à la carte » au client et qui se superposeront au dictionnaire générique si l'utilisateur le souhaite (dictionnaire de l'informatique, de la médecine, etc.).

— L'équipe SEMTAB : Nous avons vu que l'équipe du dictionnaire entrait une seule traduction pour chaque mot. Mais prenons par exemple le verbe anglais « to take » : équivalent dans le dictionnaire = « prendre ». Or, considérons les deux cas suivants :
Let's take a brake = Faisons une pause
He took advantage of him = Il a abusé de lui

Dans ces deux exemples « to take » n'est pas traduit par prendre. Comment faire, donc, pour obliger la machine à utiliser la traduction souhaitée ? C'est là tout le rôle de SEMTAB : elle repère et codifie tous les cas où le nom, le verbe, etc., ne prennent pas la traduction du dictionnaire.

Conservons l'exemple de « to take » : l'équipe a répertorié 138 traductions différentes dans des contextes variés. Voilà comment elle aurait réagi dans le cas du deuxième exemple : He took advantage of him. D'abord elle aurait codifié « took » puis « advantage ». Elle aurait codifié ensuite « abuser » et lui aurait ajouté une suite de chiffres signifiant que quand on rencontre « to take » en anglais suivi d'assez près par « advantage » (il peut exister un adverbe de quantité entre les deux mots) la

machine doit choisir le verbe « abuser » en français et annuler tout équivalent pour « advantage ». Les choses se compliquent quand il y a effectivement un adverbe de quantité entre les deux mots : imaginez qu'il y ait dans le texte « He took too much advantage of him » : « Too much » qui est amalgamé à « advantage » en anglais disparaît avec lui dans la traduction française. Cette difficulté est peut-être résolue à l'heure actuelle mais il est possible qu'elle fasse encore partie des fautes à corriger par le traducteur qui utilisera le système et qui obtiendra « Il a abusé trop de lui ».

Quand une phrase arrive dans le système, celui-ci repère d'abord les éléments principaux de la phrase et les considère dans l'environnement le plus large. Dès qu'il reconnaît des combinaisons d'éléments composant une règle spécifique de SEMTAB, il s'arrête et retient la traduction donnée.

S'il ne reconnaît aucun élément de l'entourage du mot en question, il donne la traduction du dictionnaire.

A mon départ, fin 1985, l'équipe SEMTAB avait écrit 8 000 règles environ. Le niveau de raffinement de la traduction a été évalué par IBM qui, à cette époque, devait prendre en charge la suite de l'élaboration du système, après la compagnie franco-américaine Burroughs qui l'avait financé jusque-là, et qui estimait le produit satisfaisant pour ses besoins en traduction.

Après de nombreux tests à partir de documents non édités (manuels d'utilisation, lettres, etc.) IBM a conclu que le système lui économiserait, en l'état, environ 80 % du temps de traduction, laissant 20 % de « post-editing » [*] à son équipe de traducteurs. Ils ont acheté le système tel quel. Pour que l'efficacité du système s'améliore de 5 points (passe à 85 %) il aurait fallu y travailler encore de très nombreux mois vu le niveau de subtilité où il était arrivé.

La rapidité de l'élaboration du système anglais-français est due en grande partie à une collaboration très étroite avec l'équipe qui avait élaboré le système anglais-allemand.

Technologie LOGOS pour une traduction automatique

Le processus d'acquisition linguistique tel qu'il a été défini par Chomsky sous-tend la technique d'établissement de différents systèmes de traduction automatique.

Les principes de la grammaire générative, ou transformationnelle, développés par Chomsky ont déterminé des modèles linguistiques par lesquels peut être générée la composition syntaxique des phrases en accord avec un ensemble de règles de transformation. Ce processus de

génération de phrases part d'éléments syntaxiques tels que les verbes, les noms, etc.

Cependant Chomsky lui-même reconnaît que les « modèles » linguistiques en question ne reflètent pas le processus linguistique complet qu'utilise l'esprit humain pour se servir du langage. Il est le premier à avouer que les procédés mentaux réels ne sont pas encore cernés. Les modèles de création de phrases semblent inspirés par la recherche en informatique menée au Massachusetts Institute of Technology (MIT) à Boston. Cette idée est renforcée par le fait que les « modèles » sont facilement adaptables à un système de traduction automatique alors qu'ils posent des problèmes insurmontables aux humains.

Si l'on enseigne une langue selon ces modèles on s'aperçoit que les étudiants acquièrent un langage artificiel sans réelle adéquation avec le fonctionnement de l'esprit humain.

Autre aspect de la théorie de Chomsky qui pose un problème, si l'on veut suivre ses préceptes dans l'apprentissage d'une langue ou l'élaboration d'un système de traduction automatique : il sépare la syntaxe de la sémantique, prétendant que le sens d'un élément donné de la langue existe en dehors de la syntaxe et que la syntaxe d'une structure de phrase en tant que telle n'a pas de portée sémantique. Il semble donc bien que la grammaire générative ait peu contribué à éclaircir le problème principal de l'élaboration d'un système de traduction automatique qui est *d'apprendre à une machine à simuler le processus linguistique humain.*

Le système LOGOS écarte les principes de grammaire transformationnelle et essaie d'expliquer l'acquisition du processus linguistique humain par la même occasion.

Prenons des exemples :

Exemple n° 1 :
John wants Marie to study music
John veut que Marie étudie la musique

Exemple n° 2 :
John asks Marie to study music
John demande à Marie qu'elle étudie la musique
ou : John demande à Marie d'étudier la musique

En anglais, les deux constructions de phrases semblent identiques alors qu'en français on a des structures complètement différentes. Dans un cas pareil, lorsqu'on enseigne une langue, on reporte la difficulté à plus tard, au moment où l'étudiant aura atteint un niveau suffisant pour comprendre. La question est de savoir ce qu'il a besoin d'apprendre avant d'appréhender ce genre de problème linguistique avec succès. Le processus est le même en ce qui concerne l'élaboration d'un système de traduction automatique.

---

* *post-editing = correction à posteriori.*

Reprenons les exemples ci-dessus :
En anglais la structure est la suivante : N1 + V1 + N2 to V2 + N3

Il est évident que dans les exemples n^{os} 1 et 2 c'est le verbe V1 qui différencie les deux phrases : en français, les verbes vouloir et demander exigent des constructions différentes.
Doit-on en conclure qu'il faille caractériser chaque verbe avant de le confier à la machine ou existe-t-il une façon plus efficace de résoudre le problème ?

Les deux phrases anglaises des exemples précités peuvent aussi être transformées à l'intérieur de la même langue :
Exemple n° 1 bis : John wants that Marie study music
Exemple n° 2 bis : John asks of Marie that the study music

Nous pouvons en conclure que « want » et « vouloir » dans les exemples 1 et 1 bis fonctionnent de la même façon à cause de leur valeur sémantique commune en français et en anglais.

De même pour les verbes « demander » et « ask » dans les exemples 2 et 2 bis. Cela pourrait vouloir dire que les structures syntaxiques ne sont pas de purs faits de hasard mais qu'elles sont commandées en partie par le contenu sémantique du verbe.

Cette interférence est renforcée par d'autres observations : dans les exemples 1 bis et 2 bis, les verbes « want » et « ask » ne sont plus interchangeables. Cela implique l'existence d'une forte relation entre valeur sémantique et comportement syntaxique.

D'autre part, si l'on remplace les verbes en question par d'autres verbes au fonctionnement syntaxique semblable on s'aperçoit qu'il existe entre eux une certaine similitude sémantique. Exemple : vouloir, exiger, réclamer...

La conclusion de ces remarques est qu'on ne peut séparer syntaxe et sémantique comme Chomsky voulait le démontrer (structure profonde et structure de surface). On doit plutôt considérer un ensemble de phénomènes que l'on désignera sous le nom de propriétés sémanto-syntaxiques.

On peut poursuivre la comparaison entre l'élaboration d'un système de traduction automatique et le processus d'apprentissage d'une langue étrangère :

Un professeur n'explique pas tout d'abord, en théorie, l'utilisation des prépositions. Il commence par les employer dans différents contextes et l'étudiant se forme petit à petit un système cohérent (la base même de tout processus d'apprentissage). Il comprend peu à peu que telle préposition associée à tel verbe a telle valeur sémantique et que la même préposition associée à un autre verbe a une tout autre valeur.

Exemples :
John blocks the flow with a valve
John bloque l'écoulement *au moyen* d'une valve / *avec* une valve

John acquaints the man with the facts
John met l'homme au courant *des* faits

John does not mix with the students
John ne se mêle pas *aux* étudiants

Un étudiant intelligent réalise que les changements dans la traduction ont quelque chose à voir avec la signification du verbe. Par exemple, si le verbe « bloquer » est suivi par un élément précédé d'une préposition il s'agit sans doute d'un instrument quelconque qui permet de réaliser l'action de blocage. L'étudiant comprend presque inconsciemment que la nécessité de l'emploi d'une préposition vient du fait que le verbe n'indique pas par quel moyen on va « bloquer » quelque chose. Les prépositions « au moyen de », « à l'aide de », lui viennent immédiatement à l'esprit.

Qu'est-ce que tout cela signifie ?

Un étudiant n'a pas besoin d'apprendre 10 000 verbes et leurs constructions propres pour maîtriser la langue française. Il suffit de lui donner 30 à 50 types de constructions avec un verbe représentatif de chacune d'entre elles. Quand il rencontre un nouveau verbe, il le compare aux modèles connus. Quand il a déterminé la catégorie de verbes à laquelle il peut l'assimiler, il lui attribue le fonctionnement-type de cette catégorie. Il peut se tromper, bien entendu, mais c'est aussi un moyen de perfectionner ses connaissances. Apprendre à déceler les traits sémanto-syntaxiques plus subtils de cette façon assure à l'étudiant une progression très rapide dans son apprentissage de la langue et des moyens très efficaces pour appréhender une deuxième ou une troisième langue étrangère.

Qu'est-ce que tout cela implique pour la traduction automatique ?

On peut faire apprendre une langue à un ordinateur de la même façon qu'on la fait acquérir à un étudiant doué. Il s'agit de lui présenter les mots avec leurs propriétés sémantiques aussi bien que syntaxiques. Pour cela on attribue à chaque mot une série de codes numériques puisque le langage de la machine est un langage chiffré.

L'ordinateur ne pense pas comme un humain mais il a une formidable mémoire, pratiquement infaillible.

Dans la mémoire du système LOGOS, on a stocké des « règles linguistiques » qui consistent en combinaisons de codes sémanto-syntaxiques et qui permettent au système de résoudre un grand nombre d'ambiguïtés au niveau sémantique, fournissant la base d'une traduction automatique de bonne qualité.

Je ne peux entrer plus avant dans les détails de fonctionnement du système LOGOS pour des raisons de confidentialité, mais je vais essayer de résumer les grandes étapes du fonctionnement de la « machine à traduire la plus extraordinaire que j'aie jamais rencontrée ».

Le système LOGOS

Le système LOGOS s'appuie à la fois sur des principes de grammaire comparative et sur l'observation du fonctionnement inter-langues. Il traduit d'abord le langage naturel en langage codé « concentré » sémanto-

syntaxique appelé SAL (semantic abstraction language). Les techniques de la grammaire comparative servent à séparer l'analyse de la langue source de la production de la langue cible.

La principale différence entre le système LOGOS et d'autres systèmes comme EUROTRA est que, dans le système LOGOS, quelques aspects de la production de la langue cible apparaissent en parallèle à l'analyse de la langue source, plutôt qu'à la suite de celle-ci.

Le fonctionnement automatique du système se déroule en huit étapes :

Etape 1

Procédure assez mécanique qui consiste à convertir les textes écrits « ordinaires » en un langage que la machine puisse comprendre et, également, à repérer les formes sémanto-syntaxiques contenues dans le texte. Cette étape et l'étape n° 8 sont les seules étapes à modifier quand on veut passer d'un type de machine à un autre qui utilise des conventions de traitement de texte différentes.

Etape 2

C'est le moment où le système consulte le dictionnaire pour la première fois. Opération très importante : transformation des mots du langage naturel en langage SAL (voir plus haut). Ce que fait le système LOGOS, par cette opération, c'est simplifier la richesse sémantique du langage naturel en le considérant à un niveau plus abstrait. Le travail effectué à ce point équivaut à réduire les caractéristiques du langage naturel au centième. Cette réduction de la complexité du langage est ce qui permet au système LOGOS de ne pas être débordé quand il atteint le moment final de la traduction. SAL est un langage situé entre la syntaxe pure et la sémantique pure.

Par exemple :

Prenons le symbole syntaxique N(= nom) et le contenu sémantique du mot « chaise » : en langage SAL le mot « chaise » est remplacé par l'élément générique « surface portante » qui peut aussi recouvrir « banc », « table », étagère », etc.

Après cette étape, le système travaille exclusivement à partir d'éléments SAL. Au fur et à mesure de la progression de l'analyse de la chaîne d'éléments SAL ceux-ci se combinent et la phrase finit par ressembler à une sorte d'« épi sémanto-syntaxique ».

Au cours de cette étape on assigne aux mots et aux groupes de mots de la langue source une traduction générique dans la langue cible, traduction qui sera susceptible de modifications au cours des étapes suivantes.

Etape 3

Là, on essaie de faire disparaître toute ambiguïté en ce qui concerne le genre, le nombre, le découpage de la phrase. Ceci est réalisé en analysant les données morphologiques, syntaxiques et sémantiques de la chaîne d'éléments SAL. C'est à ce moment-là également que l'on classe certains éléments SAL comme « intouchables », certaines suites de mots comme idiomatiques et donc indissociables. Le verbe est considéré comme étant au centre de l'établissement des groupes syntaxiques.

Etape 4

Le véritable découpage de la chaîne d'éléments SAL commence. Un travail très important est exercé sur les groupes nominaux. Les éléments principaux de ces groupes sont identifiés ainsi que la plupart des éléments de leur entourage. Ils sont ensuite envoyés à SEMTAB (voir plus haut). A l'étape suivante, la chaîne d'éléments SAL sera réduite, puisque les groupes nominaux ne seront plus représentés que par leur élément principal auquel seront amalgamés tous les renseignements nécessaires à l'accord du verbe, etc.

Etape 5

Les seules parties de la phrase qui n'aient pas encore été amalgamées sont les propositions relatives et les complétives. La complexité de l'analyse de ce type de propositions (surtout en allemand/langue source) oblige le système à s'y prendre en deux fois. Le système LOGOS essaie de trouver une traduction intelligente en replaçant par exemple les éléments « pré-modifiants » en allemand en position « post-modifiante » en anglais.

Une autre activité du système à ce moment de la traduction est de décider des limites définitives des différentes propositions.

Etape 6

C'est là que le travail sémantique est le plus important. Le système a décidé à l'étape précédente quels éléments appartenaient à quelle partie de la phrase. Il va maintenant rechercher dans la banque de données sémantique quelles sont les nuances sémantiques à l'intérieur de ces différentes parties de phrases qui demanderaient éventuellement une transformation dans la langue cible en question.

Etape 7

C'est le dernier stade de l'analyse. Toute ambiguïté doit être levée en particulier en ce qui concerne le fonctionnement des différentes parties de la phrase entre elles.

Le système doit maintenant savoir exactement à quel verbe appartient quel sujet, COD ou autres compléments, adverbes, etc.

Etape 8

Cette étape finale est la synthèse, la traduction proprement dite.

Le déroulement de ces 8 étapes est *entièrement automatique.*

Si je ne parle que de « phrases » et non de « textes » c'est que le système traite chaque texte phrase par phrase. Mais les écrits que l'on soumet à la machine sont des textes authentiques, sans préparation spéciale si ce n'est la vérification des points entre les phrases et de l'orthographe normale (les fautes de frappe peuvent entraîner des erreurs d'interprétation de la part de la machine quand en anglais, par exemple, elle cherche un sujet pluriel à un verbe auquel on a simplement oublié de rajouter un « s »).

Quand j'ai quitté LOGOS, fin 1985, l'ambition de cette compagnie était de mettre au point un système « à cibles multiples » dont les premières « cibles » seraient le français et l'espagnol. L'objectif de ce projet est de permettre aux clients d'acheter non pas plusieurs systèmes indépendants mais un système de base sur lequel ils pourraient adapter de nouvelles langues cibles au fur et à mesure de leurs besoins. Langues cibles à venir : l'allemand (quelques petites modifications au système existant sont nécessaires), l'italien et le portugais.

Il est possible qu'à l'heure où j'écris cet article d'autres progrès aient été accomplis dans la recherche LOGOS en traduction automatique. Le secret professionnel, très sensible dans ce domaine, interdit de divulguer quelque information que ce soit avant qu'une année se soit écoulée depuis la réalisation d'une nouvelle découverte. Je tiens à remercier M. Bernard Scott, le président de LOGOS et M. Christopher Titford dont j'ai utilisé les notes pour faire cet article, ainsi que M. Lionel Mellet, chef de projet et tout(te)s mes collègues pour tout ce que j'ai appris avec eux dans ce domaine où nous sommes encore tous des pionniers.

Marie-Hélène Bouveret-Mounpain