

## LA TRADUCTION AUTOMATIQUE AU JAPON

(Rêve et Réalité)

*C'est là véritablement traduire, qui est de reconstituer au plus près l'effet d'une certaine cause... (Paul Valéry « Variété » ,1924).*

### 0. Préambule

Présenté pour la première fois en Europe à Nantes au 15e Colloque international de linguistique informatique (COLING-92: 20-28 juillet 1992), le système japonais de traduction automatique de la langue parlée (ATR - Automated Telephony Research Institute, Nara), était capable de reconnaître des énoncés japonais introduits grâce à un microphone et de générer leurs équivalents anglais (ou allemands) à « haute voix » grâce à un haut-parleur tout en affichant à l'écran certains résultats intermédiaires d'analyse et de synthèse. Les spectateurs pouvaient donc être témoins de la mise en correspondance (projection) de la chaîne « parlée » (qui de par sa nature phonique est linéaire) avec des structures hiérarchiques (syntaxiques et sémantiques). La distinction saussurienne entre la Langue et la Parole prend ainsi son deuxième souffle sous le vocable du « Traitement automatique de la parole » (par opposition au « Traitement automatique de la langue »). Mais cette fois, par rapport au Langage, cette distinction ne résulte pas tant d'une opposition sciemment (théoriquement) établie que de la séparation des domaines qui se sont constitués *de facto*.

### 1. Qu'est-ce que la traduction ?

Le problème principal que rencontre la théorie de la traduction est exactement le même de nos jours que dans n'importe quelle théorie linguistique: nous ne disposons toujours d'aucune théorie universelle du langage qui soit satisfaisante. Certes, étant donné que la traduction est une activité dont les buts pratiques sont essentiels nous avons tendance à lui appliquer des critères d'efficacité, ce qui n'est pas le cas de nos théories linguistiques contemporaines. Pour cette raison, il est possible de considérer la technologie de la traduction automatique (TA) comme le seul domaine dans lequel on peut se livrer à une évaluation précise de faits concernant deux langues différentes. La TA est une discipline toute jeune qui n'en est encore qu'à l'époque des pionniers. Cependant, tout comme dans les autres domaines de la linguistique informatique, la recherche en TA a déjà permis d'aller plus loin sur certains points et même de découvrir plusieurs nouveaux problèmes. C'est pourquoi la TA n'est pas

tout simplement une “science des automates” même si son objectif est de construire une machine capable de traduire d'une langue à l'autre.

Au risque d'enfoncer une porte ouverte, il n'est pas inutile de rappeler que la qualité des traductions dépend largement de l'habileté et du savoir du traducteur. Les mauvaises traductions sont en effet l'œuvre de mauvais traducteurs. C'est pourquoi on peut considérer que la traduction est une opération de projection des constructions d'une langue sur celles d'une autre langue, opération qui fait appel à des procédures complémentaires (contextuelles, pragmatiques et cognitives). La traduction est donc une conversion opérant sur des discours linguistiques et faisant usage d'un savoir sur le monde. Ainsi, traduire veut dire faire coïncider deux discours, rapprocher deux cultures, c'est-à-dire: multiplier les différences.

## 2. Aperçu historique des recherches

A l'heure actuelle, on qualifie souvent de génération 0 les premières tentatives de traduction automatisée. La raison en est que les systèmes développés alors ne tenaient compte que des rapports d'équivalence “mot-à-mot” au cours du processus traduisant. Au cours de la période des recherches ultérieures, qui a été appelée 1<sup>ère</sup> génération, les systèmes de traduction tiennent déjà compte de règles syntaxiques rudimentaires mais ce sont seulement les recherches récentes (celles de la 2<sup>e</sup> génération) qui se fondent sur le traitement du sens.

Au Japon, le premier système de TAO s'appelait YAMATO. Ce système, réalisé en 1958, possédait 600 transistors et 7000 diodes ainsi qu'une mémoire de 100 Ko de DRAM, ce qui était énorme à l'époque. Plusieurs programmes de traduction de l'anglais vers le japonais ont été implémentés sur Yamato entre 1959-1961. Entre 1964-68, suivant des technologies entièrement repensées, trois universités (Tokyo, Kyoto et Kyushu) se sont lancées dans la recherche des systèmes à traduire.

Mais c'est seulement depuis 1978 qu'il est devenu possible de traiter les kanji. En effet, c'est à cette époque là que la société Toshiba a commercialisé le premier traitement de texte japonais que l'on connaît aujourd'hui comme *wâpuro* (abréviation de “word processor”). Dorénavant, on a vu apparaître de plus en plus de recherches dans le domaine du traitement automatique de la langue japonaise car la conversion des kana/romaji en kanji a nécessitait l'analyse morphologique des syntagmes (et non uniquement des plus petites unités du discours). Le développement technologique ultérieur passait par l'augmentation de la puissance de calcul et de la mémoire de masse.

Le premier projet japonais significatif dans le domaine de TA s'appelait «Mu-Project» (1982-86) et faisait partie d'une série de projets concernant les recherches visant la conception des ordinateurs capables de comprendre le langage humain qui ont démarré aux alentours de 1983, date à laquelle le gouvernement japonais avait annoncé l'intention de soutenir ce genre de recherches sous le bien ambitieux intitulé «Ordinateurs de 5<sup>e</sup> Génération» (c'est-à-dire: ordinateurs intelligents). Il s'agissait donc du premier projet gouvernemental (Secrétariat d'État à la Science et Technologie) de cette envergure dans le domaine de l'intelligence artificielle au Japon. Bien qu'il soit connu sous le nom de “Projet de la cinquième génération d'ordinateurs” (Dai-go-sedai Computer Project), son nom officiel fut en fait “Recherches sur le système japono-anglais de transmission rapide des informations contenues dans la documentation scientifique et technique” (Nichi-ei kagaku-gijutsu bunken no sokuhô shisutemu ni kan-suru kenkyû).

«Mu-Project», visant la traduction entre le japonais et l'anglais (Nakamura J., 1988), avait pour objectif de créer: (a) une *base de données*: dictionnaire de terminologie scientifique, (b) un *logiciel*: dispositif de conversion des structures syntaxiques, (c) un *système de traduction automatique* combinant les deux. En pratique, pour des raisons de commodité (familiarité des développeurs du système), le domaine d'application a été réduit aux questions relatives à la technologie électronique.

D'une manière générale, les problèmes informatiques ne se laissent pas aisément séparer des problèmes linguistiques sinon que ces derniers sont pris presque entièrement (sic) en charge par les informaticiens. Certaines stratégies que les ingénieurs du langage naturel doivent définir lors de la construction des systèmes de TA, par exemple, celles qui ont été adoptées dans le cadre de Mu-Project ont été présentées dans les Travaux de Linguistique Japonaise Ishiwata T. & Włodarczyk A., vol. 10, 1997).

A partir de 1993, on commencera à développer au Japon des recherches concernant la compréhension et la génération des dialogues s'appuyant sur le traitement intégré du son, de la langue et des concepts. Dix importantes entreprises d'électricité, d'électronique et de communication ainsi que plusieurs sociétés éditrices de logiciels s'étaient engagées à développer des systèmes de TA. Aujourd'hui, après la première période de commercialisation des produits développés, la plupart de ces activités de recherche se sont transformées pour pouvoir intégrer le traitement de l'information multilingue avec pour objectif de construire des outils pour la navigation sur les futures autoroutes d'information telles que Internet.

### **3. Méthodes de traduction automatique**

Les espoirs des pionniers de la TA se sont soldés par un échec car ils étaient fondés sur l'incompréhension totale des problèmes linguistiques. SYSTRAN (Georgetown University, États-Unis), premier système de TA connu aujourd'hui comme système de «génération 0» consistait notamment tout simplement en la traduction “mot-à-mot”. (Notons cependant qu'à l'heure actuelle, la version bien évoluée de ce système est utilisée expérimentalement par la NASA et par EURATOM.) L'introduction des problèmes syntaxiques n'a pas amélioré les résultats des recherches malgré les espoirs que nourrissaient les informaticiens des années 1970.

#### **3.1. Représentation sémantique universelle (interlangue)**

Les tentatives d'utiliser une représentation sémantique universelle comme méthode pour traduire automatiquement étaient moins naïves mais - tout comme les précédentes - elles se fondaient sur l'hypothèse de la faisabilité à courte échéance. Au Japon, les protagonistes de la méthode de la *traduction intégrale* voyaient en l'apparition de logiques plurivalentes étendues ou encore en la représentation universelle du sens (réseaux sémantiques de Ross Quillian - 1968 et représentation conceptuelle de Roger Schank 1975) les outils nécessaires et suffisants pour atteindre la haute qualité de la traduction. Le langage sémantique intermédiaire appelé “langage pivot” ou “interlangue” était devenu l'objet de recherches en TAO des années 1980. Mais la problématique de grammaire “universelle” ou “particulière” s'étant fait sentir, les laboratoires japonais ont vite adopté la méthode utilisant le principe de “transfert des structures intermédiaires” car, comme l'a observé Nagao M. (1983), “les langues contiennent plus d'exceptions que de cas qui pourraient être traités avec des règles clairement définies, d'une part, et présentent beaucoup de domaines qu'on ne peut pas expliciter logiquement, de l'autre”.

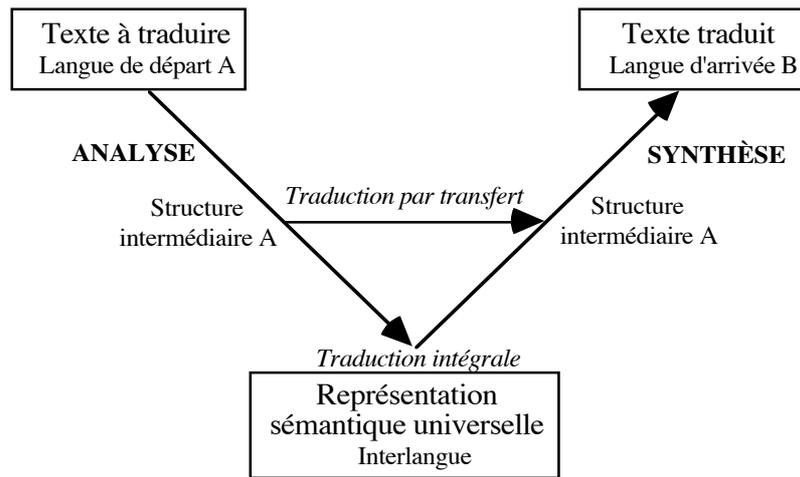


Fig. 1: Représentation schématique des premières approches en traduction automatique

Ainsi, suivant le degré de profondeur d'analyse sémantique des textes, vers 1980 on distinguait entre les trois types d'approches suivants: (a) traduction "mot-à-mot", (b) traduction "par transfert" au niveau intermédiaire, (c) "traduction intégrale" à représentation sémantique universelle.

### 3.2. Transfert des structures intermédiaires

Il va sans dire que la méthode dite "par transfert" n'était qu'un compromis entre la méthode "mot-à-mot" et la méthode "intégrale" et ne se justifiait guère que par des raisons économiques dans la mesure où, même si les systèmes de représentation sémantique universelle ne nécessitaient pas de programmation particulière pour chaque langue, l'élaboration de tels systèmes constituait (et constitue toujours) un enjeu très coûteux. Précisons encore que ce compromis concerne la profondeur de l'analyse sémantique, de là le nom de "transfert au niveau intermédiaire" donné à cette méthode.

On peut dégager deux techniques à l'intérieur de la méthode "par transfert" selon le format de la représentation des énoncés issue de l'analyse:

- (1) transfert des arborescences syntagmatiques (ATLAS/I, KATE, ATHENE etc.),
- (2) transfert des arborescences casuelles (Mu-PROJECT, ATLAS/II, VENUS, TRAP, LUTE, TAURAS etc.).

Le transfert se définit donc comme conversion des structures intermédiaires de la langue de départ (source) en structures intermédiaires de la langue d'arrivée (cible). Le transfert ne peut concerner que des structures analogues: structures de constituances ou de dépendances respectivement.

### 3.3. Recombinaison des schémas

La méthode (aujourd'hui, il est déjà possible de parler "des méthodes" au pluriel), qui utilise les schémas des phrases et des syntagmes alignés de deux (ou plusieurs) langues, constitue l'alternative aux approches qui se fondent sur les règles de réécriture.

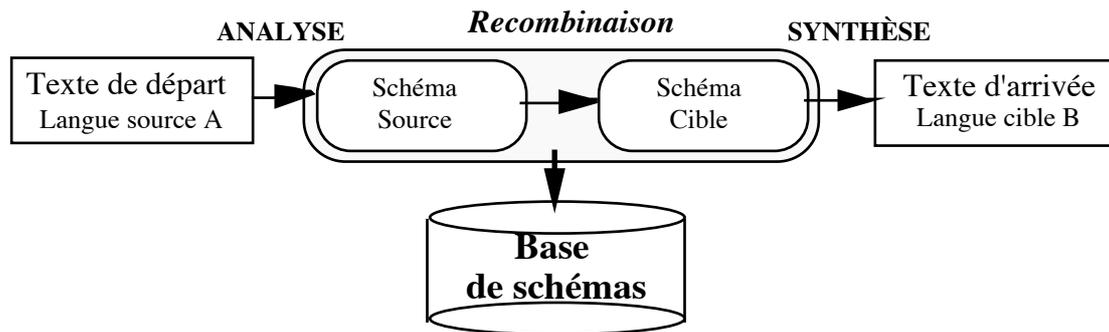


Fig. 2: Traduction automatique utilisant des schémas alignés

Les schémas ici représentent les phrases et les syntagmes sont alignés dans une base de schémas. La traduction consiste donc à trouver dans la base le schéma dont la structure est la plus proche de la phrase à traduire (analysée) et à générer une phrase dans la langue cible tout en recombinant les éléments définis dans les deux schémas alignés. Ce type d'approche donne de résultats très satisfaisants surtout quand on doit traduire les expressions figées. Pour cette raison, on dit aussi qu'il s'agit là des méthodes de TA "par recherche des analogies". Il va sans dire que la base des schémas analogues doit être préalablement établie et cela pose beaucoup de problèmes car pour être suffisante, elle doit être bien volumineuse.

### 3.4. Transfert conceptuel (systèmes à base de connaissances)

Les systèmes de TA à "transfert conceptuel" (c-à-d.: systèmes à base de connaissances) utilisent trois types de connaissances: (a) *grammaticales*, (b) *lexicales*, (c) *spécifiques* aux domaines d'application.

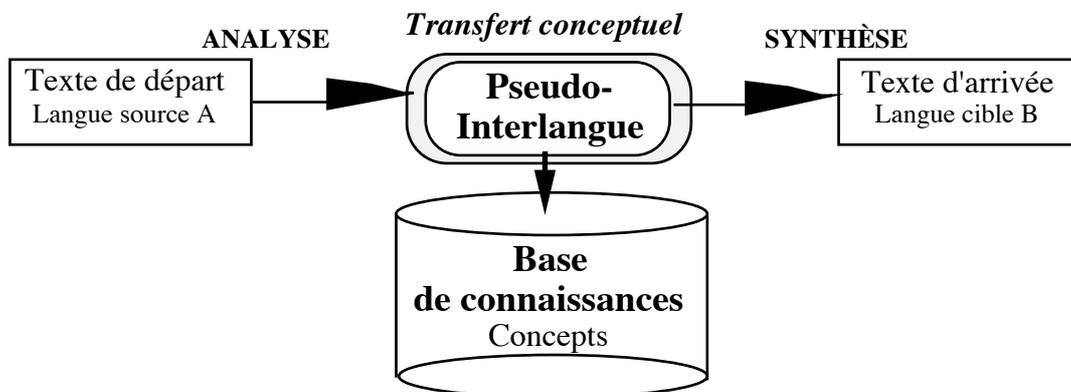


Fig. 3: TA utilisant les connaissances spécifiques au domaine d'application

La TA basée sur les connaissances dont le projet représentatif au Japon est SHALT2, IBM-Japon (1995- ), outre les connaissances relatives à la grammaire et aux dictionnaires, utilise en effet les définitions des concepts, les concepts, les descriptions des projections (syntaxiques et conceptuelles) ainsi que les règles de transfert de concepts. Tout comme dans le cas de la TA par transfert, ici également toutes ces connaissances sont exploitées par les trois programmes suivants: analyseur, module de transfert, synthétiseur. Dans SHALT2, le processus de traduction se déroule de la façon suivante: d'abord, le texte de départ est converti en une représentation conceptuelle. Ensuite, cette représentation est transférée à une

représentation conceptuelle propre à la langue d'arrivée. Enfin, la génération se déroule en sens inverse de l'analyse.

En résumant, selon le genre de connaissances exploitées lors du processus de TA, on distingue aujourd'hui entre (1) la traduction utilisant les règles grammaticales et les dictionnaires (2) la traduction utilisant les exemples bilingues, (3) la traduction utilisant les connaissances symboliques et/ou statistiques et (4) la traduction utilisant l'extension bilingue de la conversion des kana en kanji.

#### **4. Évaluation**

Du point de vue linguistique, il est intéressant d'étudier quelles structures (syntaxiques et sémantiques) des phrases japonaises seraient plus appropriées pour servir d'échantillons dans les tests d'évaluation pendant la phase initiale. Il est important de souligner que dans le domaine de la TA aujourd'hui les critères de ce genre d'expérimentations vont beaucoup plus loin que ceux proposés par l'école générativiste pour les études contrastives.

Le Comité d'évaluation des phrases de l'Association Asie-Pacifique pour la Traduction Automatique (AAMT) s'est justement occupé d'évaluer les "phrases japonaises difficiles à traduire automatiquement". Actuellement, le Comité étudie les signes de ponctuation dans les phrases anglaises et japonaises. Les résultats sont publiés au fur et à mesure dans les revues SIG NL, IPSJ, et la revue de l'AAMT (Yokoyama S. 1994).

#### **5. Conclusions**

La période de recherches très intenses en TA des langues naturelles dans les années 1960 aux États-Unis s'est soldée par la déclaration de l'infaisabilité à courte échéance faute de ressources informatiques adéquates (le rapport ALPAC- 1966). Ce défi a été relevé au Japon par le lancement spectaculaire en 1982 du projet appelé "la 5e génération d'ordinateurs" (système informatique doté d'intelligence artificielle et équipé d'un ensemble de connaissances parmi lesquelles les capacités linguistiques). Plus d'une trentaine de projets japonais de recherche en TA ont vu le jour depuis cette date. Comme il fallait s'y attendre, tous se sont révélés vains. Faut-il conclure que la TA restera à jamais impossible ? Pour répondre à cette question, il faut tout d'abord rappeler que le savoir traductologique (art de traduire) dépend des sciences du langage (explication des phénomènes linguistiques). De même, la traductique (technique informatique de traduire) dépend des deux dernières. Les recherches continuent dans tous ces domaines.

Dans une perspective moins ambitieuse que celle à laquelle nous a habitué à la traductologie (avant tout la traduction des oeuvres littéraires), il est tout à fait raisonnable d'espérer que les avancées en linguistique et en informatique permettront à l'avenir de construire une sorte d'intelligence artificielle suffisamment puissante pour traduire d'une langue à l'autre. Certains chercheurs ont déjà décidé de repenser les principes mêmes des ordinateurs actuels tout en mettant à profit les connaissances interdisciplinaires des domaines tels que la neurobiologie, les neurosciences, l'intelligence artificielle (voire computationnelle), la science cognitive et la linguistique. Car, afin de traduire comme les humains, il faut apprendre à faire comme les humains. Matsuo Bashô a dit au 17e siècle: "ce qui concerne le pin apprend-le du pin, ce qui concerne le bambou, apprend-le du bambou".

André WLODARCZYK

## 6. Bibliographie

- ISHIWATA Toshio & WLODARCZYK André (1997) “Le Japonais et l’ordinateur - présentation générale”, dans «*Travaux de linguistique japonaise*» (publication de l’Université de Paris 7), volume 10 - “Langue - Ordinateur - Mentalité”, Paris : 7-27
- MIZUTANI Shizuo (1991) “Description Systématique de la Grammaire Japonaise”, trad. par Reiko Shimamori et André Wlodarczyk (pp. 15-175), suivi de “Application de la grammaire de Mizutani au traitement informatique” par André Wlodarczyk, (pp. 177-198), dans «*Travaux de linguistique japonaise*», Vol. IX, Université de Paris VII, Paris
- NAGAO Makoto (1983) “Machine Translation Project of the Japanese Government”, Atelier de traduction automatique Europe-Japon, Commission de la Communauté Européenne, Luxembourg
- NAKAMURA Jun-ichi (1988) *A Software System for Machine Translation*, Doctoral Thesis, Kyoto University : 183 pp.
- YOKOYAMA Shôichi (1994) “Honyaku-konnan-reibun no kikai-honyaku to hyôka [Traduction automatique des phrases difficiles à traduire et évaluation des résultats]”, in *Shizen gengo shori* 101-6, Tokyo, pp 41-48