

TRADUCTION AUTOMATIQUE

UTOPIE OU REALITE ?

F. KADI ALLAH,

Département de

traduction,

Es-Senia, Oran.

Si la traduction dite humaine (ou entièrement humaine) est le moyen le plus fiable pour 'traduire' les textes, il n'empêche que la lenteur la caractérisant demeure son talent d'Achille dans un nouveau contexte où la 'ruée' vers l'information, due essentiellement à la prolifération d'informations nouvelles de plus en plus nombreuses et de plus en plus variées, dans des langues qu'on maîtrise peu ou pas, est quotidienne. Le recours à la traduction automatique est devenu ainsi inévitable pour 'résorber' cette gigantesque masse d'informations. Que l'on juge : les grosses compagnies telles que IBM, SONY traduisent 1 milliard p/an, la commission européenne 1.200.000 p/an, la traduction sur le web est quant à elle beaucoup plus énorme.....c'est donc cet espoir pour accéder instantanément à l'information dans les autres langues et pour ses enjeux économiques que l'intérêt pour la T.A va en grandissant. Un responsable d'une mission française écrivait déjà en 1984 : ' la langue est un élément primordial de la production industrielle : une centrale nucléaire représente 100 000 pages de documentation technique. Une automobile, un central téléphonique, un ordinateur, sont autant de produits qu'il n'est pas question de produire, réparer, exporter sans faire un important effort de rédaction et de traduction technique dont le coût représente un pourcentage important du prix total du produit... '.

La T.A ou traductique est l'usage de l'informatique dans le processus de traduction dans le but d'automatiser le transfert linguistique d'un texte, pris comme 'échantillon' ou représentant, de la langue source vers la langue cible. La T.A ne comprend bien

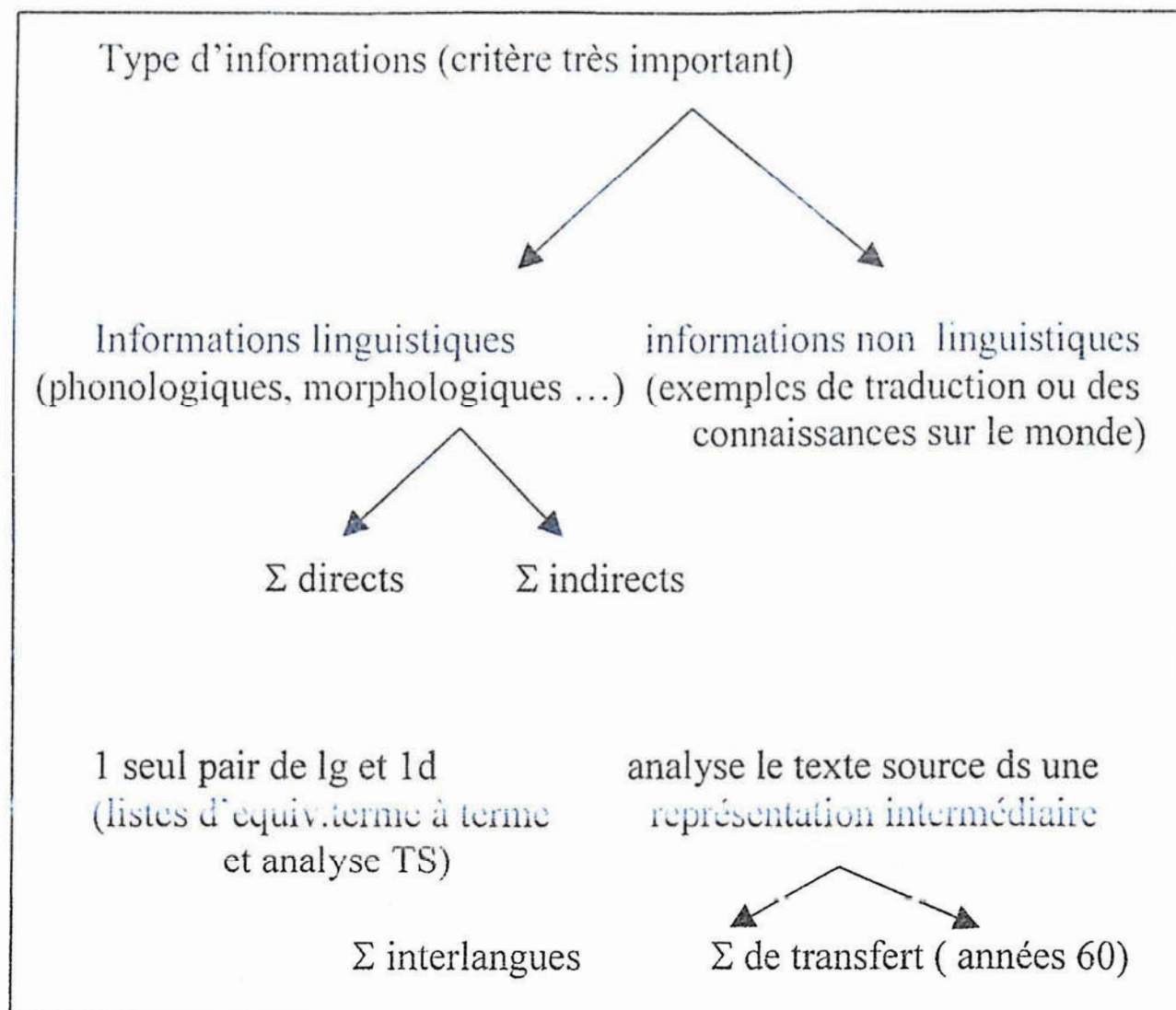
la langue source vers la langue cible. La T.A ne comprend bien évidemment pas la T.H.A.O (traduction humaine assistée par ordinateur) et peut être scindée en deux grands domaines qui sont d'une part la T.A.A.H (traduction automatique assistée par l'homme) ou T.A.O (traduction assistée par ordinateur) ou encore T.A.A (traduction automatique assistée), et d'autre part la T.E.A.H.Q (traduction entièrement automatique de haute qualité) qui reste un objectif bien éloigné.

Cet intérêt pour la T.A ne date pas d'aujourd'hui. C'est ainsi que les premiers traducteurs automatiques ont vu le jour en 1946 date qui coïncide avec la première calculatrice électronique (!). Les premiers systèmes reposaient sur la traduction mot à mot en procédant en 3 étapes : la lemmatisation, le transfert lexical de la langue source vers la langue cible et enfin un réajustement morpho-syntaxique. Si ce procédé simpliste prenait en charge certains aspects de la langue tel que l'aspect paradigmatique, il ne rendait toutefois pas compte de sa complexité. Et c'est tout naturellement qu'on s'aperçut vite de ses limites. Ainsi on ne pouvait traduire mot à mot les mots composés, les mots polysémiques ... sans altérer le sens et aboutir à des déconvenues criardes. L'outil de traduction automatique ou ordinateur avait par conséquent besoin de comprendre. Qu'est ce que ça veut dire donc comprendre pour un ordinateur ? c'est tout simplement, et tout à la fois conceptualiser, modéliser et représenter : c'est à dire faire abstraction de la forme au détriment du contenu sémantique. La forme ou graphie n'étant utile que pour le traitement différentiel ou analogique entre les unités sémantiques pour l'extraction du sens 'Bien comprendre' est donc lié à la capacité de l'ordinateur de calculer une représentation sémantique du texte source.

La modélisation nécessite le recours à deux types d'information (voir schéma ci-dessous) selon qu'on soit rationaliste ou empiriste. Les rationalistes prônent le recours aux informations dites linguistiques (phonologiques, morphologiques ...). Les empiristes quant à eux préfèrent les informations dites non linguistiques comme les exemples de traduction ou des connaissances sur le monde. Techniquement, il existe deux types de systèmes de T.A : les

Traduction Automatique Utopie ou Réalité ?

systèmes directs qui traitent une seule paire de langues utilisant des listes d'équivalence terme à terme avec au préalable une analyse du texte source, et les systèmes indirects qui procèdent par une analyse du texte source dans une représentation intermédiaire. Ces derniers peuvent être soit des systèmes à interlangue (représentation du sens dans un langage neutre) ou des systèmes de transfert procédant en trois étapes : analyse, transfert et génération.



représentation du sens dans 1 lgge neutre
(LS → interlangue → LC)

3 étapes
(analyse – transfert –

génération)

Il serait alors légitime de se poser la question sur le degré d'efficacité de tels traducteurs électroniques. Pour répondre à une telle question on peut dire que l'efficacité est subordonnée ou

tributaire à leurs aptitudes à surmonter les difficultés inhérentes à la T.A. Les difficultés ont alors deux origines. Ceux qui découlent des informations non linguistiques comme les difficultés probabilistiques ou les difficultés d'enregistrement de tous les exemples possibles et dont il n'est pas question dans cet article. Et les difficultés engendrées par les informations linguistiques c'est à dire des difficultés rencontrées en analyse et en génération partant du fait que la langue est une infinité d'ambiguïtés. Ainsi d'une part une forme linguistique dispose de plusieurs sens possibles (difficultés en analyse) et d'autre part un sens donné peut se présenter sous différentes formes linguistiques (difficultés en génération). Ce sont des difficultés lexicales. En plus l'interaction de ces formes linguistiques entre elles engendre ce qu'on appelle les ambiguïtés syntaxiques des lors que le recouvrement de leur champs sémantiques n'est pas toujours évident. On se contentera de donner quelques notions de difficultés d'ordre lexical accompagnées d'exemples pris indistinctement du Français ou de l'Anglais.

Les difficultés lexicales :

Elles découlent de la conception particulière du 'mot' par l'ordinateur puisqu'il est perçu comme une suite de caractères encadrée de séparateurs (espace blanc ou signe de ponctuation). Sachant qu'il existe dans des langues comme le français ou l'anglais par exemple des expressions composées telles que pomme de terre, rouge-gorge, pied-à-terre, hard-ass, hard copy ..., il devient difficile à l'ordinateur de distinguer ces mots composés comme 'mot' unique. L'unique solution reside au recours à la traduction dite compositionnelle c'est à dire répertorier tous les mots composés et leurs equivalents dans la langue cible, chose qui n'est pas evidente quand on sait qu'il en existe deja environ 2 millions (français et anglais) et la prolifération de nouveaux autres exclusivement techniques tels que floppy disk, virus-ridden...et la liste n'est pas close... !!! Il y a aussi dans le meme registre les expressions verbales composées comme porter aux nus, crier fort, to feel hard done ...et toutes les formes idiomatiques anglaises. La première tâche à effectuer donc est de répertorier ou stocker toutes les expressions composées et savoir reconnaître les expressions verbales dans les

Traduction Automatique Utopie ou Réalité ?

phrases parce qu'elles ne sont pas forcément contiguës comme le montre l'exemple suivant :

- Il fut porté aux nues.
- Tous les présents ont porté cette vedette aux nues.

On trouve aussi des difficultés homographiques :

- de catégories de discours différents (nom, verbe, préposition...)
- le garçon (article défini) ; je le vois (pronom)
- Il marche rapidement (verbe) ; la marche est le meilleur sport (nom)
- de même catégorie mais de sens différent (polysémie):
 - Il mange un avocat (fruit) ; Il a rencontré un avocat (homme de loi)
 - la marche est programmée pour demain (manifestation) ; la marche est bénéfique à la santé (sport) ; tu montes la marche (batiment).
 - his eye hurts him (organe) ; to have a good eye (flair) ; the eye of the needle (couture).

Ce sont là quelques ambiguïtés lexicales introduites par l'ordinateur dont l'identification sémantique n'est nullement perceptible à moins d'utiliser d'algorithmes ou méthodes de raisonnement basées sur des règles linguistiques (grammaire ...) ou sur des connaissances du monde.

En guise d'illustration d'algorithmes utilisant les règles linguistiques pour lever l'ambiguïté lorsqu'il s'agit de catégories de discours différents:

Il marche rapidement
(il) (marche) (rapidement)
il=pronom, marche=nom
l'ordinateur se pose la question
marche=nom/verbe ?
levée de l'ambiguïté
il=pronom, pronom.nom interdite, pronom.verbe autorisé
Donc marche=verbe

En ce qui concerne les ambiguïtés de même catégorie mais de sens différent, le meilleur moyen étant d'utiliser la catégorisation c'est à dire la classification sémantique de toutes les catégories grammaticales. Ensuite on définit toutes les relations conditionnelles entre ces catégories afin de 'detecter' ou 'délimiter' le champ sémantique du mot. Ainsi si on considère les deux exemples suivants :

- l'élève saisit la leçon.
- L'élève saisit le stylo.

L'ordinateur ne peut trancher entre les deux homographes 'saisit' dans le sens de 'comprendre' dans le premier et 'saisit' dans le sens de 'tenir' dans le second pour pouvoir opter pour la traduction appropriée de ces deux mots. La procédure de délimitation du champ sémantique serait par exemple comme suit :

Classification sémantique des verbes :

Saisir1 = classe de 'action mentale'

Saisir2 = classe de 'action physique'

Classification ou catégorisation des compléments du verbe :

Leçon = classe 'abstrait'

Stylo = classe 'concret'

Relation conditionnelle :

Dire que pour 'saisir1' c'est à dire appartenant à la classe de 'action mentale' le complément d'objet doit appartenir à la classe 'abstrait', sinon c'est à dire si le complément appartient à la classe 'concret' l'ordinateur optera pour 'saisir2' c'est à dire classe 'action physique'.

Donc pour la levée de l'ambiguïté lexicale concernant la polysémie, il faut procéder par catégorie et ensuite faire une interaction fonctionnelle et conditionnelle entre les catégories. Le problème qui se pose dans ce cas c'est l'intersection entre les catégories ou classes et la prolifération des combinaisons si on tient compte du fait qu'il peut exister des 'interférences' entre elles en plus de la multitude inimaginable de classes qu'on peut concevoir. Les relations conditionnelles génèrent des graphes et arborescences qui restent difficile à gérer.

Traduction Automatique Utopie ou Réalité ?

On peut citer d'autres ambiguïtés comme les ambiguïtés croisées et les ambiguïtés pragmatiques Elles feront le sujet d'un prochain article.

Conclusion :

Bien qu'on n'ait pas passé en revue toutes les difficultés auxquelles l'ordinateur se heurte, et elles sont nombreuses, le lecteur pourra déduire de lui-même que la gestion de tous les problèmes de la langue par la machine lors du processus de traduction (lexicaux, syntaxiques, morphologiques...) est une mission délicate et actuellement impossible à réaliser. La raison majeure est que la traduction automatique fige le mot, lui dénie son aspect expressif et suggestif presque malgré elle. L'automatisation de la traduction est normative par l'ensemble des règles qu'elle prescrit sous forme procédurale ou algorithmique.