Vers une nouvelle époque en traduction automatique

John Hutchins
(University of East Anglia, Norwich, England)

Troisièmes Journées Scientifiques LTT, Montréal, 30. septembre 1993

1. Introduction

Environ 1989 la traduction automatique a commencé une période d'innovation méthodologique qui a changé l'optique de recherche.

Qu'est-ce qui a changé? Quelle était la situation dans la traduction automatique il y a cinq ans? A partir de 1975 jusqu'à 1988 on a vu apparaître un grand nombre de systèmes opérationnels et commerciaux: Systran, Logos, Météo, ainsi que plusieurs systèmes japonais, et autres. Ils sont en général basés ou bien sur la méthode "directe" de traduction, ou sur la méthode de transfert syntaxique. Ils se reposent sur des dictionnaires bilingues assez riches pour les domaines des textes à traduire; l'analyse linguistique n'est pas très profonde ou abstraite, il n'y a presqu'aucune analyse sémantique, et l'exploitation des connaissances non-linguistiques est tout à fait absente.

Quant à la recherche, on peut dire sans crainte de contradiction que, jusqu'à la fin des années quatre-vingt le cadre dominant a été l'approche basée sur les règles linguistiques: les règles d'analyse syntaxique, les règles lexicales, les règles pour la formation de représentations abstraites, les règles de désambiguïsation, les règles de transformation des arbres syntaxiques, les règles de transfert lexical, les règles de génération syntaxiques et morphologiques, etc. A partir d'environ 1985, on a commencé àdévelopper des systèmes basés sur les connaissances du domaine des textes àtraduire, mais cette approche est restée une nouveauté presque jusqu'à la fin de la décennie.

Depuis 1989 ce cadre dominant a été rompu par l'entrée en scène de méthodes et de stratégies nouvelles qu'on appelle maintenant les méthodes "basées sur corpus". En premier lieu, un groupe à IBM a publié en 1989 les résultats de ses expériments avec un système de traduction purement statistique. L'efficacité de cette méthode a surpris beaucoup de chercheurs et a inspiré l'expérimentation avec les méthodes statistiques dans les années suivantes. En second lieu, certains groupes japonais ont commencé en même temps précisement à publier leurs résultats préliminaires avec des méthodes basées sur un corpus d'exemples de traductions. La caractéristique principale des deux approches est que les textes sont comparés et les traductions lexicales sont choisies sans utiliser aucunes règles syntaxiques ou sémantiques.

Dans cette conférence je vais concentrer sur les nouveaux développements de la recherche et je ne décrirai pas aucun projet en détail; les systèmes cités sont seulement des exemples; il existe beaucoup d'autres (pour les citations voir mon étude récente (Hutchins 1993). Je ne dirai également presque rien au sujet des méthodes déjà bien établies à la fin des années quatre vingt. En outre, je ne vais pas parler de l'utilisation des systèmes commerciaux, ni des aides automatisées pour traducteurs. Mon sujet exclusif est le développement de méthodes nouvelles dans la recherche de traduction automatique. Bien entendu, beaucoup des méthodes sont expérimentales et n'ont pas été mis à l'épreuve dans les systèmes à grande échelle. Toutefois, les tendances que je décrirai sont réelles; la traduction automatique a subi un renouveau de sa méthodologie dans les récentes années.

2. Systèmes basés sur des règles

Avant de décrire ces nouveaux développements je commencerai par les approches basées sur des règles parce que il y a eu ici également des développements théoriques et méthologiques d'une assez grande importance.

Il y a cinq ou six ans on a vu la fin de deux grands projets de l'approche "transfert": le projet Ariane à l'Université de Grenoble et le projet Eurotra des Communautés Européennes. Ces systèmes illustraient les traits typiques des systèmes dits de la "deuxième génération", c'est-à-dire: trois étapes d'analyse, de transfert et de synthèse; des processus d'analyse et de génération qui parcourent plusieurs niveaux distincts de morphologie, de syntaxe et de sémantique; des représentations d'interface assez abstraites en forme d'arbres étiquettés; l'utilisation des règles pour transformer des arbres d'une étape à l'autre; traitement par lots avec post-édition et sans aucune intervention humaine au cours de la traduction; et une absence presque totale d'information pragmatique et textuelle.

Cependant, les systèmes de transfert ont continué. Il y a, par exemple, le système commercial Metal, le projet LMT d'une équipe à IBM, et le projet multilingue Eurolang qui est en train d'être développé par la compagnie SITE en France avec la collaboration de la compagnie allemande Siemens-Nixdorf, et qui se base sur les expériences du projet Eurotra.

L'approche "interlingue" a continué, elle aussi, avec encore plus d'énergie. Les traits distinctifs de cette approche sont bien connus: une langue pivot neutre pour représenter le sens des textes (l'interlingue,) et des banques de connaissances dans le domaine des textes à traduire. A l'Université de Carnegie Mellon plusieurs modèles ont été développés, et en 1992 on annonca l'inauguration d'un projet collaboratif avec la compagnie Caterpillar, dont le but sera un système pour la traduction sans post-édition des manuels techniques dans le domaine des engins de terrassement.

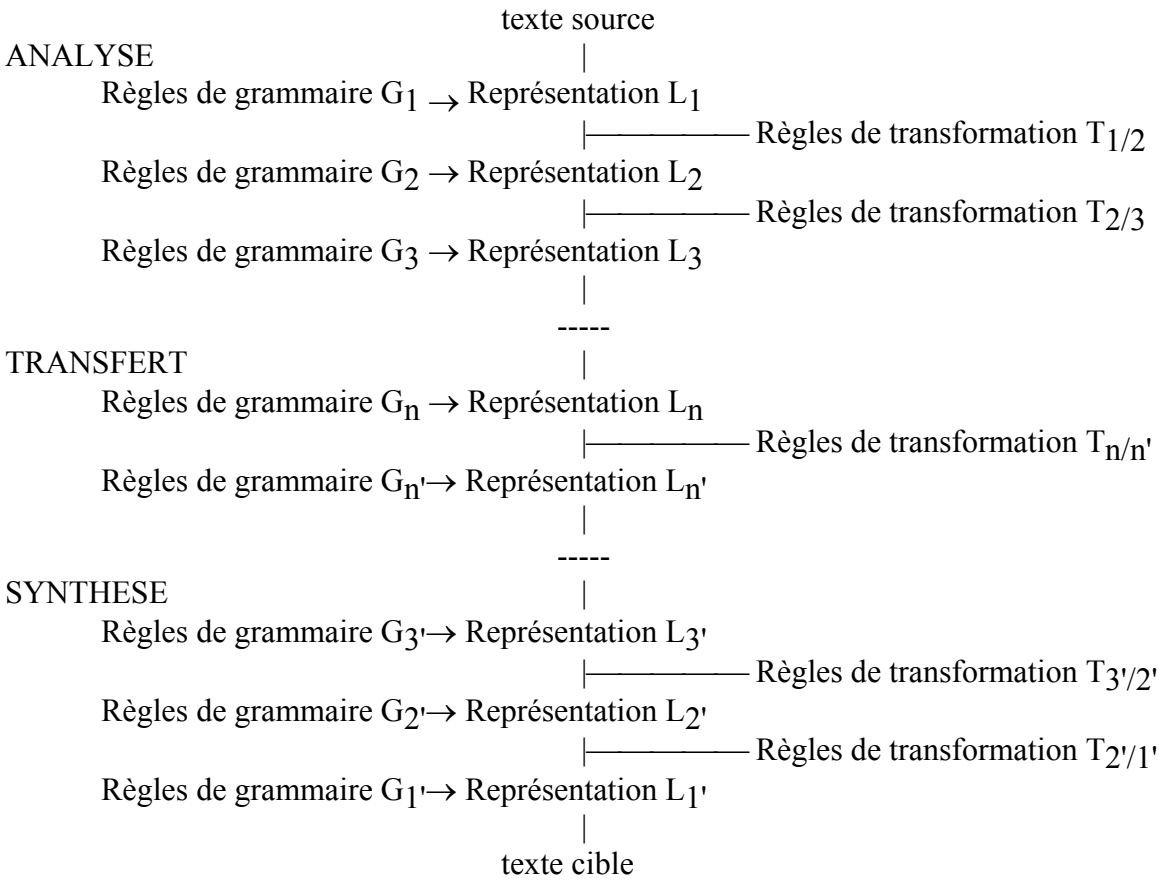
Il existe d'autres systèmes "interlingue", par exemple le système ULTRA à l'Université de New Mexico et le système UNITRAN basé sur la théorie linguistique des principes et des paramètres. On doit y ajouter le projet Pangloss, un système interlingue limité au sous-langage des fusionnements et des rachats. Celui-ci est un projet collaboratif des universités de Southern California, New Mexico State et Carnegie Mellon, et qui utilise les expériences de ces équipes dans leurs propres recherches.

3. L'entrée de la tendence "lexicaliste"

Une trait caractéristique des systèmes basés sur les règles a été la transformation ou 'mappage' des représentations en forme d'arbres étiquettés. Par exemple *[fig.1]*, dans le système Eurotra on a proposé une série de transformations des arbres: un arbre morphologique est transformé en un arbre syntaxique, un arbre syntaxique en un arbre sémantique, un arbre d'interface du texte source en un arbre équivalent du texte cible, etc. Essentiellement, un arbre doit satisfaire des conditions précises, posséder une structure particulière et contenir des unités lexicales particulières ou des traits syntactiques ou sémantiques particuliers. En plus, les arbres eux-mêmes sont testés par les règles de formation - une grammaire met à l'épreuve la structure et les relations qui y sont représentées. Un arbre est rejeté s'il ne conforme pas aux règles grammaticales du niveau en question: morphologie, syntaxe, sémantique, etc. Les grammaires

et les règles de transformation déterminent les conditions ou contraintes qui limitent les possibilités de transfert d'un niveau à un autre et, en somme, d'un texte de la langue source à un texte de la langue cible.

Fig.1: Règles de formation et transformation (Eurotra)



Au cours des années, on a développé des formalismes basés sur les contraintes. Au lieu d'un grand ensemble de règles qui ne sont élaborées que pour s'appliquer à des circonstances et des représentations très particulières, on a extrait un ensemble assez restreint de règles abstraites et on a transféré les conditions et les contraintes aux données lexicales. Par exemple, *[fig.2]* pour traduire le verbe anglais *like* en verbe français *plaire* il est nécessaire de transformer la structure syntaxique: le sujet anglais (*John*) devient un objet indirect en français (*à Jean*), et l'objet direct (*Mary*) devient un sujet en français (*Marie*). Ces conditions se trouvent dans les ensembles de traits morphologiques, syntactiques et sémantiques des unités lexicales elles-mêmes. Un formalisme un peu plus complexe est nécessaire pour indiquer les contraintes attachés au mot anglais *likely* et au mot français *probable*. Le mot anglais exige un complément en forme d'un infinitif, tandis que le mot français exige une phrase subordonnée.

Fig.2: Formalisme basé sur contraintes (LFG)

2 (a):

John likes Mary ↔ Marie plait à Jean

like, V:
(↑PRED) = like <SUBJ, OBJ>
(τ↑PRED FN) = plaire <SUBJ, OBJ>
(τ↑AOBJ OBJ) = (SUBJ)
(τ↑SUBJ) = (OBJ)

john, N:
(↑PRED) = john
(τ↑PRED FN) = jean

mary, N:
(↑PRED) = mary
(τ↑PRED FN) = marie

F-structure de la langue cible:

```
⌈PRED          plaire                    ⌉
|SUBJ [PRED marie ]                       |
⌊AOBJ [OBJ              [PRED jean ]   ]  ⌋
```

Student is likely to work ↔ Il est probable que l'étudiant travaillera

likely, A:
(↑PRED) = likely <XCOMP> SUBJ
(↑SUBJ) = (XCOMP SUBJ)
(τ↑PRED FN) = probable
(τ↑COMP) = (XCOMP)

probable, A:
(↑PRED) = probable <COMP>SUBJ
(↑SUBJ FORM) = il
(↑COMP COMPL) = que

F-structure de la langue cible:

```
⌈PRED          probable              ⌉
|SUBJ          [FORM          il ]   |
|               ⌈PREDtravailler ⌉     |
|COMP           |COMPL que      |     |
⌊               ⌊SUBJ [...]     ⌋     ⌋
```

Quant aux règles de transformation elles sont devenues les méchanismes informatiques d'unification. Ce sont les règles d'unification qui dirigent l'interaction des ensembles de traits, la formation de nouveaux ensembles et l'élimination des ensembles illégitimes.

Au lieu de l'orientation syntaxique qui a caractérisé beaucoup des systèmes de transfert dans le passé il existe actuellement une tendance vers les solutions lexicales. Un exemple extrême de l'approche "lexicaliste" est la méthode qu'on appelle en anglais "shake and bake" (en français peut-être "agiter et cuire") *[fig.3]*.

Fig.3 Approche lexicaliste ('Shake-and-bake')

unité lexicale monolingue (anglaise):

[12]

ORTHO	*like*		
SEM	E1:	like (E1), role (E1, experiencer, X1), role (E1, stimulus, Y1)	
ARG0E1			
ARG1X1			
ARG2	Y1		

unité lexicale monolingue (espagnole):

[13]

ORTHO	*gust-*		
SEM	E2:	gustar (E2) role (E2, stimulus, X2) role (E2, experiencer, Y2)	
ARG0E2			
ARG1X2			
ARG2Y2			

entrée lexicale bilingue pour *like-gustar*:

SPANISH	[13]	SEM	ARG0 E ARG1 X ARG2 Y
ENGLISH	[12]	SEM	ARG0 E ARG1 Y ARG2 X

Il n'y a plus aucunes représentations structurales; il n'y a que des ensembles de représentations lexicales. Le processus de traduction exige l'identification des unités lexicales de la langue cible qui satisfaisent les contraintes sémantiques qui sont liées aux équivalents lexicaux de la langue source. Une traduction est formée (ou 'cuite') par les intéractions entre les ensembles de traits et contraintes qui s'attachent aux mots de la langue cible.

Les origines des grammaires d'unification et des grammaires basées sur les contraintes remontent à presque dix ans d'ici. De nos jours, l'unification est devenu un concept central pour un grand nombre de théories linguistiques, et les grammaires et les formalismes basés sur les contraintes attirent beaucoup de chercheurs en traduction automatique: LFG (Grammaire fonctionnelle lexicale), Definite Clause Grammar, GPSG, Grammaire categoriale, etc. (Parmi les grammaires des contraintes on peut compter la grammaire d'UNITRAN basée sur les principes et les paramètres.) Le majeur avantage de ces grammaires est la simplification des règles d'analyse, de transformation et de génération. Au lieu d'une série de représentations complexes d'information à plusieurs niveaux, on voit des représentations mono-stratales ou même le transfert au moyen de unités lexicales simples. En même temps, les composantes de

ces grammaires sont en principe réversible. Il n'est pas nécessaire de construire pour la même langue des grammaires différentes d'analyse et de génération; le même formalisme et les mêmes grammaires peuvent être appliqués dans les deux sens.

Plusieurs équipes ont construit des systèmes généraux pour le traitement de la langue naturelle, basés sur les grammaires d'unification et des contraintes, qui s'appliquent aussi à la traduction. Le système CLE (Core Language Engine), par exemple, a été appliqué à la traduction automatique du suédois à l'anglais et de l'anglais au suédois; tandis que le système PLNLP (Programming Language for Natural Language Processing) a fourni la base de systèmes de traduction pour le portugais, le koréen et le japonais. Mieux connu dans le domaine de la traduction automatique est l'Environnement Linguistique d'Unification (ELU) développé par une équipe suisse à Genève. Sur cette base on a fondé un système bi-directionnel pour la traduction des bulletins d'avalanches entre l'allemand et le français.

4. Les lexiques et la génération de textes

La tendance vers les approches lexicalistes a eu deux conséquences: l'une pour les lexiques et l'autre pour la génération de textes. Pour les lexiques on observe une augmentation dans l'étendue des informations liées aux unités lexicales. Elle ne sont plus seulement des données morphologiques et grammaticales des mots de la langue-source et des mots et phrases correspondants de la language-cible. On a ajouté les contraintes syntaxiques et sémantiques et les informations non-linguistiques et conceptuelles, souvent limitées au domaine particulier des textes à traduire. Ceci est le plus évident dans les systèmes basés sur une interlingue, par exemple dans les systèmes développés à Carnegie Mellon ou dans le système UNITRAN.

On voit croître l'intérêt porté aux problèmes de la construction des lexiques pour la traduction automatique. La construction d'un lexique est difficile et coûteux s'il soit suffisamment grand pour les applications réelles et pratiques d'un systeme opérationnel. Beaucoup d'équipes cherchent des méthodes pour acquérir les informations lexicales en utilisant des sources lexicographiques déjà disponibles, par exemple des dictionnaires bilingues destinés aux étudiants ou consacrés aux différents domaines scientifiques. En même temps, les groupes de recherche collaborent de plus en plus étroitement dans la construction de lexiques embrassant une grande gamme d'applications et une grande variété de systèmes. L'exemple le plus connu d'entre eux est sans doute le projet collaboratif EDR (Electronic Dictionary Research) de plusieurs compagnies japonais.

Quant à la génération, il y a dix ans encore, on croyait que les composants qui posaient le plus de problèmes étaient ceux de l'analyse syntaxique et sémantique, de la désambiguïsation du sens, de l'identification des antecédants des pronoms - en somme, de la compréhension du texte à traduire. A cette époque on négligeait pour la plupart les problèmes de la génération de textes idiomatiques dans la langue cible. Aujourd'hui, il est évident à tout le monde que, pour améliorer la qualité des traductions automatiques il faut non seulement faire attention à l'analyse mais aussi à la sélection des mots propres et à l'idiomatisme des textes produits. La génération de textes n'est plus un aspect negligé de la recherche en traduction automatique.

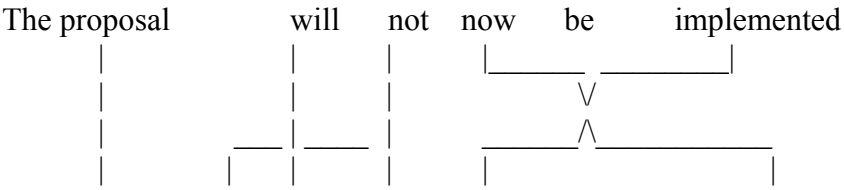
5. Systèmes basés sur un corpus de textes bilingues

Ces deux tendances ont été accentuées fortement par les développements de nouvelles méthodes qui sont basées sur les corpus de textes bilingues, en particulier les méthodes statistiques et les méthodes basées sur des exemples de traductions.

Le développement le plus étonnant a été le projet Candide par un groupe de chercheurs à IBM. Sa méthode est fondée sur l'analyse stochastique de l'énorme corpus de textes en anglais et en français fourni par les rapports des débats parlementaires au Canada. La première tâche est d'aligner les phrases, les groupes de mots et les mots individuels des textes parallèles. Ce stage est achevé sans utiliser aucune information linguistique. On voit, par exemple [*fig.4*], que cette méthode a aligné *proposal* et *proposition*, *now* et *maintenant*, et le mot anglais *implemented* et la phrase française *mises en application*. Mais, contre l'intuition linguistique elle a aligné *will* et *seront*, tandis que pour le mot *be* aucun mot n'est aligné en français. A partir d'un grand nombre de tels alignements anglais-français on a calculé les fréquences et les probablités des correspondances. Le mot anglais *not* correspond le plus souvent a deux mots français (fertilité 2 à une probabilité de 0.758), et ces deux mots sont en général *ne* et *pas* (probabilités de 0.469 et 0.460); les autres correspondances sont moins probables: *non* (0.024), *pas du tout* (0.003), etc. La **méthode a** été evaluée par la traduction de l'anglais en français.

Fig. 4: TA stochastique (Candide, IBM)

Alignement:



Anglais: *not*

Français	Probabilité	Fertilité	Probabilité
pas	.469	2	.758
ne	.460	0	.133
non	.024	1	.106
pas du tout	.003		
fau	.003		
plus	.002		
etc.			

Pour beaucoup de chercheurs, ce qui a été surprenant c'est que les résultats étaient tellement acceptables: presque la moitié des phrases traduites correspondait exactement aux traductions dans le corpus, ou bien le même sens était exprimé avec des mots un peu différents, ou bien d'autres traductions légitimes étaient offertes. Pour améliorer ces résultats, les chercheurs d'IBM ont proposé d'introduire des méthodes statistiques plus complexes et aussi quelques éléments minimaux d'information linguistique -ils sont des empiristes et ils sont prêts à

expérimenter avec d'autres moyens. Ils proposent d'utiliser l'information morphologique, p.ex. de traiter toutes les variantes d'un verbe comme un seul mot, et ils proposent aussi d'utiliser certaines transformations syntaxiques, p.ex. la transformation de structures discontinues en structures qui ressemblent plus exactement celles de la langue cible [*fig.5*]

Fig. 5: Transformations syntaxiques (proposées)

Has the store any eggs? → The store has any eggs QINV
John does not like turnips → John likes do_not_M1 turnips
Je ne sais pas → Je sais ne_pas
Je vous le donnerai → Je donnerai le_DPRO vous_IPRO

La méthode basée sur des exemples de traduction a profité, elle aussi, du développement de logiciels pour l'accès rapide aux banques de données textuelles. Pour cette méthode il y a une banque de textes bilingues en parallèle, qui ont été alignés, soit par les méthodes statistiques, soit par l'analyse morphologique et syntaxique. L'essence de la méthode est l'extraction et la sélection des phrases équivalentes. Par exemple [*fig.6.*] si on cherche une traduction pour le mot anglais *fields* on trouverait peut-être dans une banque de données les possibilités suivantes pour le français: *domaines, activités, champs*. Pour chaque exemple on trouve aussi les contextes. S'il y a une correspondance exacte (*coal fields* → *bassins-houilliers*) la sélection serait terminés immédiatement. Mais, s'il n'y a pas une correspondance exacte, il faut utiliser des algorithmes pour trouver l'équivalent propre.

Fig. 6: Banque d'exemples des traductions: *field*

anglais	français
the main fields	les principaux domaines
the following fields	les domaines suivantes
these two fields	ces deux domaines
the specialized fields	les domaines spécialisés
the para-medical fields	activités paramédicales
the magnetic fields	les champs magnétiques
the coal fields	les bassins-houilliers
the corn fields	les champs de blé

Quelques équipes utilisent les méthodes sémantiques, par exemple un réseau sémantique ou une hiérarchie de termes d'un domaine. D'autres équipes ont utilisé les informations statistiques à propos des fréquences lexicales dans la langue cible. Le plus souvent, la méthode des exemples est utilisée comme complément aux méthodes plus traditionnelles basées sur les règles linguistiques. Parce que les textes sont extraits des traductions humaines on a l'assurance que les résulats seront aussi idiomatiques que possible. C'est une des plus grandes difficultes de la traduction automatique du francais àl'anglais de choisir la préposition propre pour le petit mot *de [fig.7]*. Une banque de données qui offre un grand nombre d'exemples peut très bien aider la sélection des équivalents. Elle peut aider encore plus où il y des difficultés plus grandes, p.ex. pour la traduction en français de la phrase anglais *have an effect on [fig.8]*.

Fig. 7: *de*

français	anglais
le livre *de* mon père	my father's book
un verre *d'*eau	glass of water
il est certain *de* réussir	certain to succeed
il est capable *de* résister	capable of resisting
il vient *de* Paris	he comes from Paris
le train *de* Paris	the train to/from Paris
il partit *de* nuit	he left at night
il partit *de* bonne heure	he left in good time
je suis âgé *de* trente ans	I am thirty years old

Fig. 8: *...have a/an effect on...*

anglais	français
have a direct effect on	ont une influence directe à
have a direct effect on	intéressent directement
have a direct effect on	ont eu une répercussion directe sur
has had a marked effect on	a largement influencé
had a positive effect on	s'est avérée positive dans
had a highly negative effect on [X]	[X] en auraient été gravement affectés
will have a decisive effect on	influencera de façon déterminente
would have a detrimental effect on	aurait de fâcheuses répercussions sur

Une banque de textes bilingue en parallèle peut être également utilisée dans d'autres buts. En particulier, plusieurs équipes conduisent les expériences avec des stations de travail destinées aux traducteurs humains. L'un des groupes les plus actifs dans les méthodes statistiques pour aligner les textes bilingues (c.a.d. le groupe à AT&T Bell) prévoit l'application d'un corpus aligné comme une base de connaissances pour les traducteurs. Un autre groupe très actifs est l'équipe bien connue au Centre canadien de recherche sur l'informatisation du travail. Son but est le développement d'outils valables pour les traducteurs, y compris la facilité de chercher une banque de textes pour les exemples d'utilisation de n'importe quel mot anglais ou français dans un contexte particulier.

La méthode des exemples de textes a renforcé aussi la tendance dont j'ai fait déjà mention vers la recherche de génération de textes. En dehors de la traduction automatique, il y a des autres groupes qui s'intéressent aussi de plus en plus à la représentation des données informatiques dans une langue naturelle idiomatique. La génération de textes dans plusieurs langues a été le but de deux projets canadiens: dans le projet RAREAS c'est la production de textes anglais et français des prévisions maritimes; dans le projet LFS c'est la production de sommaires bilingues sur la marché du travail.

Un autre stimulant, qui illustre lui-même une tendance importante des cinq dernières années, c'est qu'on a reconnu les demandes pour des types de traduction qui n'ont jamais été étudié auparavant. Dans le passé, les systèmes ont été en général construits pour des gens bilingues, pour les traducteurs et pour ceux qui connaissent les langues de source et de cible. En outre, les textes traduits exigeaient la post-édition. On a negligé les besoins de ceux qui ne

connaissent pas la langue cible. Ce sont souvent des négociants et des gens d'affaires qui font le commerce à l'étranger, qui veulent communiquer une message assez simple dans une langue inconnue. Dans les années récentes quelques équipes ont expérimenté avec des systèmes où la texte à traduire est composé par une collaboration entre homme et ordinateur. Il est ainsi possible d'établir une texte que le système est capable de traduire sans se referer davantage à l'auteur, qui n'exige aucune révision et dont la qualité de traduction est bien assurée.

6. Systemes destinés aux utilisateurs particuliers

Il y a toujours eu des systèmes de traduction automatique qui sont largement limité à des domaines restraints. Même les systèmes destinés à l'usage général sont souvent effectivement limités à quelques domaines particulier. La limitation diminue les difficultes de construire un lexique suffisant, et elle diminue les problèmes d'ambiguïté et de choix des équivalents dans la langue cible. Dans ces cas, il est possible de restreindre le vocabulaire et les structures grammaticales des textes à traduire. Bien que les frais de l'édition préalable peuvent être assez élévés, la post-édition sera réduite de façon considérable. On peut aussi limiter le système lui-même à un sous-langage spécifique. Un exemple bien connu est le système Météo qui traduit, depuis quinze ans, les rapports météorologiques. Parmi les systèmes de sous-langages et de langues restreintes des années plus récentes on compte le système CRITTER pour les rapports du marché des bestiaux, les projets déjà mentionnés (ELU, Pangloss, et Caterpillar), et les projets très ambitieux destinés au développement de systèmes pour traduire la langue parlée. Le projet ATR au Japon dure déjà depuis sept ans et continuera jusqu'à la fin du siècle; c'est un système pour les renseignements sur les conférences internationales et pour les enregistrements téléphoniques dans les hôtels. Le projet VERBMOBIL en Europe vise à développer une aide portative dans les négociations commerciales face à face conduites en anglais par les Allemands ou les Japonais qui ne connaissent pas très bien la langue anglaise.

Dans le passé il n'y a que peu de systèmes construits par les utilisateurs mêmes. Un exemple connu est la PAHO (Organisation Pan-Amérique de la Santé), où deux systèmes ont été développés pour la traduction de l'anglais àl'espagnol et de l'espagnol à l'anglais. Au cours des dernières années il y a eu plusieurs systèmes construits par les utilisateurs mêmes. Ce sont typiquement des systèmes avec des vocabulaires restreints, pour un domaine particulier et basés sur un sous-langage spécifique. C'est un signe encourageant que les methodes informatiques de la traduction automatique et du traitement de la langue naturelle se sont maintenant répandus de plus en plus au delà des cercles limités de chercheurs. De tels systèmes ne sont pas innovatifs au point de vue théorique et méthodologique, mais ils sont souvent d'une construction très avancée en ce qui concerne leur technologie informatique. Je crois que c'est une tendance que nous verrons se répandre rapidement au cours des années à venir.

7. Une nouvelle époque?

A mon avis la recherche de la traduction automatique a traversé cinq époques jusqu'à nos jours. La première a commencé par le memorandum de William Weaver en 1949 qui a lancé les recherches. La seconde a commencé avec la démonstration en 1954 d'un système assez simple pour la traduction du russe à l'anglais, qui a encouragé les agences gouvernementales des Etats-Unis et d'autres pays à soutenir les expérimentations sur une grande échelle. Elle a fini avec le fameux rapport d'ALPAC. La troisième époque a duré jusqu'à environ 1975, quand on a épourvé une résurgence grâce à l'interêt croissant au Canada et en Europe. Pendant que les systèmes des deux premières époques ont été en général basés sur la méthode directe de traduction, le cadre dominant des époques depuis ALPAC était les systèmes de transfert et d'interlingue basés sur les règles. Mais, comme je l'ai déjà decrit dans cette conférence, il existe à présent de nouvelles méthodes et tendances, les approches basées sur les corpus bilingues de textes, les méthodes statistiques et les méthodes des exemples, et en outre de nouvelles méthodes basées sur les grammaires d'unification et des contraintes. Ces innovations ont apparu pendant les cinq dernières années et la recherche en traduction automatique me semble avoir commencé une nouvelle époque. Si la méthode directe a caractérisé la "première génération" et si les méthodes indirectes de transfert et d'interlingue ont caractérisé la "deuxième generation", il faut se demander quelles seront peut-être les caractéristiques de la future "troisième génération".

8. Systèmes de la "troisième generation"?

A l'avis de beaucoup des spécialistes les systèmes d'avenir combineront des méthodes de règles assez traditionelles et des méthodes statistiques et basées sur les exemples. Ils seront hybrides. Mais de quelle façon? Dans une optique possible, on peut envisager que les méthodes linguistiques des systèmes indirects fourniront la base pour l'application des méthodes basées sur les banques de connaissances, sur les données statistiques et sur les exemples de textes traduits. La base linguistique fournira les analyses syntaxiques et sémantiques assez simples, l'essentiel du transfert et l'information syntactique pour la génération. A la base on verra probablement les formalismes d'unification et les grammaires basées sur contraintes. Les autres méthodes fourniront une désambiguïsation plus souple et les informations nécessaires pour le transfert lexical et pour la production des textes idiomatiques.

Ainsi, en ce que concerne la base de règles linguistiques on peut anticiper que:

- les règles seront moins ambitieux que celles des systèmes indirects de transfert et d'interlingue;
- l'analyse syntaxique sera limitée à la reconnaissance des structures superficielles, des composantes de phrase et des rélations de dépendance;
- il n'y aura presqu'aucune analyse profonde des rélations logiques;
- l'analyse sémantique sera limitée à l'identification des rôles d'éléments dans les phrases: l'agent, l'instrument, etc.
- les informations lexicales seront extraites principalement des sources regulières comme des dictionnaires généraux destinés au grand public; par conséquence elles n'indiqueront que des catégories syntactiques et peut-être des traits sémantiques peu raffinés.
- ces traits sémantiques assez rudimentaires ne seront utilisés que pour une désambiguïsation initiale.

- les règles de transfert lexical et structural s'appliqueront peut-être à des représentations peu profondes (quoique moins simples que dans le système d'IBM).

Quant aux méthodes encore assez nouvelles:

- des exemples de traduction, stockés dans une banque de textes bilingues alignés, seront utilisés pour aider la désambiguïsation plus délicate dans l'analyse de la langue source et pour choisir les équivalents dans la langue cible
- des données statistiques sur les collocations lexicales et les fréquences du vocabulaire monolingues aideront l'analyse syntaxique et sémantique des phrases, la désambiguïsation monolingue, et le choix des phrases idiomatiques dans la langue cible
- des données sur les probabilités des équivalences bilingues seront utilisées pendant le transfert lexical
- des banques de connaissances des domaines en question aideront la désambiguïsation monolingue et interlingue

On peut anticiper aussi d'autres développements importants:

- l'emploi des méthodes de feedback pour améliorer les grammaires (ou le fond de règles) et les lexiques monolingues et bilingues
- les recherches plus approfondies sur les problèmes de discours et de style
- l'intégration de la traduction automatique dans les systèmes pour la production, la transmission et le management de documents dans les bureau (à l'exemple des postes de travail pour traducteurs)

9. L'Utilisation des systèmes

Aujourd'hui, les utilisateurs et les chercheurs sont tous réalistes. Le système de traduction complètement automatique qui produira les textes idiomatiques comparables aux traductions humaines n'est qu'une rêve pour ceux qui n'en ont pas l'expérience. On ne croit pas non plus à la réalisation pratique et économique de systèmes à l'usage général, qui traitent une grande gamme de domaines. La recherche est centré sur le développement de systèmes limités à un sous-langage ou à un domaine technique spécial. Dans les conditions propices, les systèmes loins d'être parfaits peuvent être utilisés avec profit et succès. Sans doute tout le monde aspire à des systèmes d'une meilleure qualité, mais on ne les attend pas dans un proche avenir. Il faut rappeler que c'est toujours les systèmes expérimentaux qui expérimenteront avec des techniques nouvelles. On n'attend pas l'arrivée d'aucun système commercial basé sur ces méthodes de la "troisième generation" avant la fin du siècle. Dans cette nouvelle époque les systèmes pour la marché seront basés sur les méthodes sûres, bien établies et bien éprouvées - qui toutefois ne seront pas les méthodes les plus innovatives.

Néanmoins, on doit prévoir une expansion rapide d'utilisateurs des systèmes de traduction automatique. Dans les dernières années, le nombre de pages traduites automatiquement a cru fortement - à présent plus qu'un million de pages par l'an, ou trois cent millions de mots (Vasconcellos 1993). L'expansion a eu lieu dans les grandes compagnies multinationales et dans les agences de traduction, surtout pour la traduction des manuels techniques. Mais il y a eu aussi une expansion du nombre des utilisateurs non-professionnels. Beaucoup d'eux ont acquis des systèmes bon marché pour les ordinateurs personnels - qui sont bien sûr des systèmes assez simples quant aux méthodes linguistiques et qui n'ont pas été developpés par les expertes de traduction automatique. On peut douter l'efficacité des systèmes mais les

besoins de ces utilisateurs sont incontestables. C'est un marché que peu de chercheurs spécialistes ont considéré à nos jours.

En somme, on prévoit une période d'expérimentation variée et assez incohérente quant à la théorie. En même temps on prévoit une expansion d'utilisation des systèmes de grande taille et celle du nombre d'utilisateurs de systèmes basés sur les ordinateurs personnels. En particulier, on prévoit une expansion dans le nombre d'utilisateurs par moyen de réseaux électroniques; en France et au Japon la traduction automatique a été déjà offerte sur les reseaux PC-VAN, Niftyserve et Minitel; et cette année même CompuServe annoncera bientôt une service de traduction automatique. Quelle sortes de systemes satisfairont ces nouveaux besoins? Les chercheurs devront s'adresser à ces defis, et aussi, avec d'autant plus d'urgence, à l'établissement de normes pour permettre d'évaluer et de comparer la performance, la qualité et l'efficacité des systèmes commerciaux. On attend les changements rapides dans le domaine de traduction automatique dans le proche avenir avec beaucoup d'intérêt et d'espoir.

10. Bibliographie

Les sources principales de cette conférence sont les comptes-rendus des congrès dediés à la traduction automatique: TMI-90, TMI-92, TMI-93, MT Summit III et MT Summit IV. Voir mon article de revue dans ce dernier pour les citations précises (Hutchins 1993). Pour des articles français voir aussi le numéro spécial du journal *Meta*, vol.37 no.4, décembre 1992.

TMI-90. Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, 11-13 June 1990, University of Texas, Austin, TX, USA

TMI-92. Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. Empricist vs Rational Methods in MT, June 25-27, 1992, Montréal, Canada

TMI-93. Fifth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. MT in the Next Generation, July 14-16, 1993, Kyoto, Japan.

MT Summit III, July 1-4 1991, Washington D.C., USA.

MT Summit IV: International Cooperation for Global Communication, July 20-22, 1993, Kobe, Japan.

Hutchins, W.J. (1993) Latest developments in machine translation technology. In: **MT Summit IV** (1993: 11-34)

Vasconcellos, M. (1993) The present state of machine translation usage technology, or: How do I use thee? Let me count the ways! In: **MT Summit IV** (1993: 35-46)