

La Traduction et la Technologie : un état de l'art

Margaret King

TIM/ISSCO

Ecole de Traduction et d'Interprétation
Université de Genève

Introduction

La dernière fois que j'étais appelée à dresser un bilan des technologies liées à l'art de la traduction date d'environ dix ans. En préparant ce nouveau bilan, j'ai été frappée par un paradoxe : au niveau des théories sous-jacentes aux applications informatiques touchant à la traduction, je ne peux pas identifier ni de grands progrès, ni de changements révolutionnaires, mais force est de constater des changements énormes dans la vie quotidienne de nombreux traducteurs et dans la formation des traducteurs. L'article actuel cherche à explorer ce paradoxe.

L'état de l'art il y a dix ans

Il y a dix ans, la technologie reconnue en linguistique informatique était basée sur l'élaboration de jeux de règles censés décrire le comportement morphologique, syntactique ou sémantique de la langue à traiter. L'informatique imposait, comme toujours, ses exigences en termes de clarté, de précision et de traitement de détails sur de tels jeux de règles ; leur création reste inévitablement un travail fastidieux et de longue haleine. La linguistique également connaissait ses limites : même s'il était possible de décrire presque complètement le processus de formation des mots, au moins pour les langues de grande diffusion, il n'y avait pas – et il n'y a toujours pas – une description complète de la syntaxe d'aucune langue. Au niveau de la sémantique, l'ambiguïté omniprésente dans la langue semblait poser un problème quasi insurmontable : bien que quelques systèmes aient contourné le problème en faisant le choix astucieux de rester dans un domaine de discours limité, aucune solution générale ne semblait ni satisfaisante du point de vue théorique ni réalisable du point de vue pratique.

Cette « barrière sémantique » a été très largement discutée dans la littérature : nous n'irons pas beaucoup plus loin ici, sauf pour nous rappeler de quelques exemples qui illustreront l'ampleur du problème. Si je vous dis que je vais « acheter un canapé », comment

allez vous décider si j'ai faim ou si j'ai besoin d'un nouveau meuble ? Le logiciel qui devrait proposer une traduction de *canapé* vers l'anglais doit quand même choisir entre *sandwich* et *sofa*. Donc, le logiciel aussi doit être doté d'un moyen de savoir si j'ai besoin de la nourriture ou des meubles. Comment lui fournir ces informations ? Si je vous dis que *j'ai vu la dame avec les jumelles*, comment allez vous savoir si la dame est accompagnée par des enfants jumelles, ou si j'ai utilisé un instrument optique quand j'ai vu la dame ? De nouveau, la traduction change selon le choix fait : *I saw the woman with the twins* ou *I saw the woman through the binoculars*. L'interlocuteur humain, en règle générale, sait quel choix faire parce qu'il a suivi toute une conversation avant d'arriver là, ou il y a d'autres éléments non-linguistiques qui l'éclairent. Comment faire pour rendre disponible à un logiciel la même connaissance du contexte particulier ? Le gros du problème réside dans cette dernière phrase : si l'on regarde de près, pratiquement chaque phrase qu'on utilise contient des mots ambigus, ou est construite d'une façon qui permet plus qu'une seule interprétation du vouloir dire. L'être humain puise dans toute son expérience chaque fois qu'il participe à l'activité de communication par la langue jusqu'au point où, la plupart du temps, il n'est même pas conscient des ambiguïtés potentielles. L'ordinateur ne vit pas : le logiciel n'a pas d'expérience qui lui permet de participer intelligemment à la communication. Et si l'on voulait qu'il en fasse semblant, il faudrait lui expliquer toutes les expériences possibles et tous les contextes possibles, exprimés en plus dans un langage formel dénué de toute possibilité de nuance ou d'interprétation multiple. Il n'est pas très surprenant que personne n'ait encore relevé le défi.

Il est parfois possible, comme nous avons déjà remarqué en passant, d'éviter le problème en limitant soit le domaine de discours soit la richesse d'expression de la langue utilisée, soit les deux. Ainsi, si on sait qu'on est dans le domaine des meubles, il est beaucoup plus probable que le canapé sera un *sofa* qu'un *sandwich*. Si l'utilisation des syntagmes qui commencent par une préposition est interdite, on ne peut même pas dire *J'ai vu la dame avec les jumelles*, et les paraphrases telles que *J'ai vu la dame. J'ai vu les jumelles* ou *J'ai vu la dame. La dame a des jumelles* ou *J'utilisais les jumelles. J'ai vu la dame* enlèvent l'ambiguïté.

Des applications qui travaillent à l'intérieur d'un seul domaine et qui limitent volontairement la richesse d'expression permise existaient déjà il y a dix ans. D'ailleurs, une des premières réussites de la traduction automatique, le système TAUM-Météo (Isabelle, 1987) était dans ce style et date de bien avant, ayant produit quotidiennement depuis 1976 des traductions depuis l'anglais vers le français des bulletins météo au Canada. Moins connu, mais

encore plus vieux, le système TITUS en France utilisait un langage très réduit au niveau syntactique et ne traitait que le domaine de l'ingénierie de textiles (Ananiadou, 1987).

Les systèmes tels que TITUS et TAUM-Météo indiquaient un chemin potentiellement très fructueux, et en effet, l'utilisation de la traduction automatique combinée avec un domaine spécifique et un langage réduit s'est beaucoup développée ces derniers temps. La documentation technique dans des domaines comme l'industrie de l'automobile ou l'aéronautique est prise en charge en grande quantité par cette stratégie, sous l'appellation « langage contrôlé », et il y a eu depuis 1996 une série de manifestations scientifiques dédiées au sujet (voir <http://www.controlled-language.org>, où les actes de ces congrès sont disponibles).

D'autres utilisations de la traduction automatique il y a dix ans concernaient essentiellement des applications où une qualité de traduction relativement médiocre permettait quand même l'achèvement d'une tâche basée sur l'exploitation de la traduction. Par exemple, si on veut trier un ensemble de documents selon leur sujet principal, ou si on veut tout simplement saisir l'essentiel d'un document sans trop aller dans les détails, une traduction même mauvaise peut mener à un résultat satisfaisant. (Pour d'autres idées sur les utilisations possibles des mauvaises traductions voir Church et Hovy, 1993 – un article assez important dans la mesure où il aidait à casser l'idée qu'une traduction imparfaite ne pouvait jamais servir à rien).

C'est dans ce contexte qu'on voit pour la première fois jusqu'à quel point la Toile a influencé le monde de la traduction automatique dans les derniers dix ans. On s'y est tellement habitué maintenant qu'on a tendance à oublier qu'un accès grand public à la Toile est un développement relativement récent. Le premier serveur public est devenu disponible en 1991, mais le public était toujours un public assez réduit, composé principalement de spécialistes. Ce n'était qu'en 1993, avec la création et mise à disposition publique des premiers protocoles qui permettaient une utilisation aisée à travers une interface graphique conçue pour un utilisateur non-spécialiste (Mosaic et un peu plus tard Netscape) que la Toile d'aujourd'hui a commencé à se créer. (Pour des informations supplémentaires, voir, par exemple, Abrams, 1998).

Au tout début, la Toile était presque entièrement anglophone, mais, de plus en plus, les utilisateurs demandent un accès à la Toile dans leur propre langue ainsi que la possibilité d'accéder aux informations dans les langues autres que la leur. (Voir Lebert, 1999 pour une discussion de l'expansion du français sur la Toile.) La présentation de Charles Wayne sur le programme TIDES aux Etats Unis illustre le point : une personne monolingue, sans aide informatique ou humaine pour les langues qu'elle ne connaît pas, peut exploiter une

proportion minimale de toutes les informations disponibles, et cette proportion devient encore plus minuscule chaque année, puisque le nombre de langues utilisées sur la Toile et la quantité des informations publiées ne cessent pas d'accroître. (Wayne, 2002)

L'utilisation de langues autres que l'anglais sur la Toile est à l'origine de ce qu'on pourrait presque appeler une vulgarisation de la traduction automatique. Plusieurs systèmes sont disponibles gratuitement sur la Toile, et, le plus souvent, quand on fait la recherche des documents, la présentation des résultats offre la possibilité de demander une traduction automatique de chaque document retrouvé. Et il est intéressant de noter que même si les logiciels utilisés dans la production des traductions sont essentiellement basés sur la même technologie d'il y a dix ans, à savoir l'utilisation de règles linguistiques, la qualité des traductions obtenue peut être beaucoup plus satisfaisante qu'avec les systèmes de l'époque. L'explication réside dans une remarque que nous avons déjà faite : la construction de systèmes basés sur les règles linguistiques est un travail fastidieux et de longue haleine. Dans les dix ans qui se sont écoulés, les constructeurs des systèmes ont eu le temps de créer des jeux de règles plus complets et, surtout, d'alimenter les dictionnaires qui sont la clé de voûte de tout système « linguistique » de traduction automatique.

La même remarque s'applique également aux applications linguistiques moins ambitieuses que la traduction automatique. Il y a dix ans, il y avait déjà des vérificateurs d'orthographe et de grammaires, mais ceux qui utilisaient l'informatique dans la création de leurs documents ne s'en servaient guère. En ce qui concerne les vérificateurs d'orthographe, le problème était souvent un problème de convivialité, et le développement d'interfaces plus conviviales ainsi que l'intégration de l'outil dans les traitements de textes ont résolu les problèmes. Mais les vérificateurs de grammaire souffraient à la fois de lacunes – ils ne trouvaient pas de fautes mêmes parfois frappantes – et d'une espèce d'arrogance souvent agaçante – ils critiquaient des phrases tout à fait correctes ou des tournures comme la voix passive ou des phrases qu'ils trouvaient boiteuses parce que trop longues. L'écoulement du temps a permis au moins de combler certaines lacunes. Les vérificateurs de grammaire de nos jours identifient une bonne partie des fautes d'accord ou de syntaxe, et leur utilité est indiscutable, surtout pour celui qui rédige dans une langue qui n'est pas la sienne.

Les technologies en voie de développement il y a dix ans.

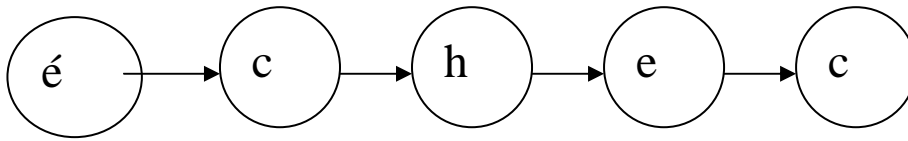
Le calcul de probabilités

Jusqu'ici nous avons fait un bref survol des technologies déjà courantes il y a dix ans, et nous avons vu comment elles se sont développées entre temps. Nous nous retournons maintenant vers les technologies qui étaient pour ainsi dire les technologies en voie de développement. Parmi celles-ci, peut être celle qui a eu des conséquences les plus profondes au niveau de la vie quotidienne des traducteurs est l'élaboration de systèmes basés sur un calcul de probabilités.

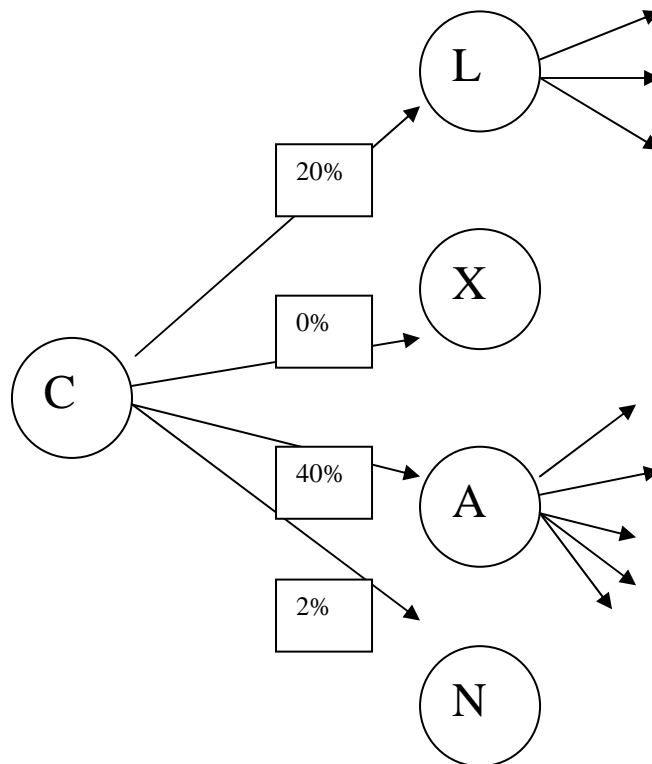
Avant d'aller plus loin, il est peut être important de remarquer que l'idée de prendre un grand ensemble de textes (un *corpus*), de l'analyser afin de déceler des comportements systématiques dans l'utilisation de la langue et enfin d'utiliser les résultats de cette analyse dans moult applications informatiques n'était pas nouveau en soi. Déjà, dans les tout premiers travaux sur la traduction automatique, on trouve des propositions dans ce sens et même des fortes controverses entre les "empiricistes", qui voulaient tout baser sur l'étude de corpora et les "rationalistes", qui voulaient modéliser le comportement humain à travers des jeux de règles. Mais les ordinateurs des années 50 et 60 n'avaient ni la puissance de calcul ni la capacité de mémoire nécessaire au développement de systèmes empiriques. En plus, il n'y avait pas de grands corpora disponibles sous forme électronique, condition préalable sine qua non pour l'existence même de tels systèmes. Vers la fin des années 80 et le début des années 90, la machine à taper a pratiquement disparu en faveur de l'utilisation des systèmes de traitement de textes sur ordinateur. Ainsi, il devient beaucoup plus facile de trouver des documents électroniques aptes à une exploitation par des systèmes informatiques. (Pour les applications multilingues, le problème de trouver des textes parallèles, où l'un des textes est la traduction de l'autre (souvent appelé un *bi-texte*) a perduré un peu plus longtemps – ce qui explique pourquoi les Actes du Parlement canadien, un des rares bi-textes disponibles relativement tôt, ont servi comme base pour plusieurs systèmes pionniers.

Puisque cette technologie est assez récente, il vaut la peine de consacrer un peu de temps à une explication de comment elle fonctionne. (Je remercie mes collègues Bruno Cartoni et Marianne Starlander pour les exemples utilisés ci-dessous.)

Un point de départ relativement intuitif est l'idée qu'on peut regarder un mot comme une suite de caractères:



Les langues différentes utilisent des caractères différents, avec une fréquence différente: par exemple, l'anglais n'utilise pas les caractères accentués qui sont typiques du français, mais même pour les caractères communs aux deux, le français utilise le *p* et le *u* beaucoup plus que l'anglais, et l'anglais utilise le *w* et le *y* beaucoup plus le français. Quand on regarde une suite de deux caractères, les différentes probabilités se distinguent encore plus nettement: par exemple, si on commence avec le caractère *c* en français, il est beaucoup plus probable que le caractère qui suit soit *l* ou *a* que *x* ou *n*. On peut différencier plus finement pour arriver à un schéma de probabilités comme le suivant, où la probabilité qu'un caractère donné suit le premier *c* est exprimé en termes de pourcentage (les chiffres sont intuitifs, et ne sont pas basés sur un corpus spécifique).



L'étude de corpus permet de calculer les probabilités de séquences de deux caractères (des *bigrammes* dans le jargon des études empiriques) relatives au corpus en question. Si le corpus est très grand et constitué d'une façon que toute une langue est bien représentée (au lieu de se contenter d'une partie seulement, par exemple, la langue de la documentation technique ou la langue des lycéens) on peut considérer les probabilités valables pour cette langue en général. Et si on peut calculer les probabilités pour des suites de deux caractères, on peut également calculer les probabilités des suites de trois (les *trigrammes*) ou de plus. Voici les trigrammes les plus fréquents pour quelques langues européennes, calculés sur un corpus de 100'000 caractères par langue:

Allemand	ich	der	die	ein	sch	nde	che	den	end	ung
Anglais	the	ion	ent	and	tio	ing	ati	tha	nth	hat
Danois	for	der	den	det	nde	ere	ing	ede	til	ter
Espagnol	ent	que	del	ión	nte	est	ela	con	res	sde
Français	ent	ion	que	tio	ons	les	men	des	ati	ela
Italien	ion	one	ent	del	zio	ell	che	con	lla	nte
Néerlandais	nde	van	and	end	het	ver	ing	een	oor	gen
Portugais	ent	que	nte	com	men	est	ode	con	ade	res

Une fois qu'on a compris les principes de base, l'utilisation des probabilités dans les applications pratiques perd un peu son air de magie. A titre d'exemple, nous prendrons une application relativement familière, la composition de messages sur un téléphone portable. Le problème de base réside dans le fait que le clavier du téléphone n'a que douze touches, bien que l'alphabet du français comporte 26 caractères. Comment peut-on taper 26 lettres quand il n'y a que douze touches à disposition ? La solution est de lier la lettre voulue à un certain nombre de pressions sur la touche. Ainsi, si on pousse une fois sur la touche 2, le résultat est la lettre *a*, si on pousse deux fois le résultat est la lettre *b*, trois fois la lettre *c*. Ainsi, le nombre de caractères possibles avec un clavier très limité est multiplié. Mais composer un message avec ce clavier réduit n'est pas toujours facile. La touche est un peu sensible – le résultat produit est la lettre *b* au lieu du *c* voulu : il faut faire marche arrière et re-essayer. Le processus de composition devient délicat et fastidieux, et les utilisateurs par paresse ou par frustration renoncent à l'utilisation de la messagerie. Les constructeurs alors ont intégré à la messagerie un petit logiciel qui facilite l'entrée des mots, en utilisant les probabilités pour opérer un choix entre les trois ou quatre caractères possibles à chaque pression. Avec ce

système, celui qui compose un message fait une seule pression sur la touche chaque fois. Le logiciel « devine » parmi les lettres associées à la touche la lettre voulue en fonction de la probabilité relative des caractères en fonction des lettres déjà choisies. Voyons, par exemple, ce qu'il arrive quand l'utilisateur veut écrire le mot "chaise".

1. Le mot commence avec un *c*, donc il appui sur la touche 2, à laquelle sont associés *a*, *b* et *c*. Il voit apparaître sur l'écran le caractère *a*, puisque la lettre *a* est plus fréquent en français que le *b* ou le *c*.
2. La deuxième lettre de « chaise » est *h*, donc il appui sur la touche 4, à laquelle sont associées les lettres *g*, *h* et *i*. Il voit apparaître sur l'écran la séquence *ai*. Le *i* est choisi parce qu'*ai* est plus fréquent en français que toutes les autres séquences qu'on peut produire en français avec *a*, *b* ou *c* comme première lettre, *g*, *h* ou *i* comme deuxième lettre (par exemple *ag*, *bh*, *cg*, et, bien sur, *ch*).
3. La troisième lettre de « chaise » est de nouveau *a*, donc il appui encore une fois sur la touche 2. Il voit sur l'écran *cia*, la séquence de trois lettres la plus fréquente parmi celles qui ont *a*, *b* ou *c* comme première lettre, *g*, *h* ou *i* comme deuxième lettre et *a*, *b* ou *c* comme troisième lettre. Notons que maintenant qu'il a une séquence de trois, le logiciel a changé d'avis sur la première lettre : les séquences qui commencent avec *a* sont moins fréquentes que la séquence *cia*.
4. La quatrième lettre de « chaise » est *i*, donc l'utilisateur appui sur la touche 4 de nouveau. Il voit apparaître la séquence *chah*, la plus probable de toutes les séquences formées d'*a*, *b* ou *c* suivi de *g*, *h* ou *i*, suivi de *a*, *b* ou *c*, suivi de *g*, *h* ou *i*.
5. La cinquième lettre, *s*, est parmi les lettres associées à la touche 7 (les autres sont *p*, *q* et *r*). Le résultat un peu surprenant est le mot *biais*, la plus probable des suites possibles.
6. Et finalement, l'utilisateur appui sur la touche 3, pour la lettre finale de « chaise », *e*. Et la séquence affichée devient *chaise*, le mot voulu. Néanmoins, *chaise* n'est pas la seule solution possible. Si l'utilisateur demande au logiciel de voir les autres possibilités, *chaire* et *biaisé* sont affichés. Le lecteur peut vérifier que les deux obéissent aux contraintes imposés par l'ordre de touches utilisées et par les lettres associées à chaque touche. Mais *chaire* est moins fréquent en français que *chaise*, et *biaisé* moins fréquent que *chaire*.

Dans l'explication, nous avons travaillé avec la suite de caractères qui constitue un mot, mais il est également possible d'appliquer le même principe aux sons de la langue, aux suites de mots et même, dans une moindre limite, aux suites de phrases. Beaucoup d'applications de ce principe de base sont arrivées sur le marché pendant les dix ans qui nous concernent. Il y a même des systèmes de traduction automatique basés sur le principe qu'on

peut calculer la probabilité relative qu'un mot ou une phrase dans la langue source sera traduit par un tel ou tel mot ou par une phrase donnée de la langue cible.

Les systèmes de dictée.

Parmi ces applications, les logiciels qui traitent la langue parlée sont peut être l'application qu'on remarque le plus. Pour prendre de nouveau l'exemple du téléphone portable, il est courant maintenant d'incorporer la possibilité de demander une communication tout simplement en prononçant le nom de la personne avec qui on veut parler au lieu de taper le numéro de la personne sur le clavier. Et on commence tous à être habitué aux systèmes de renseignements qui utilisent soit la reconnaissance de la voix soit la génération d'un message parlé.

Dans cette gamme de produits informatiques, les logiciels qui permettent la dictée d'un texte revêtent d'un intérêt particulier pour les traducteurs. Il y a presque une ironie; avant l'installation un peu partout de systèmes de traitement de texte, la pratique quasi universelle chez les traducteurs était d'enregistrer une version dictée de leurs traductions, qui était ensuite dactylographiée par des dactylographes professionnels. L'informatisation du processus de production des documents a fait que les traducteurs aussi se sont mis à maîtriser les outils informatiques: il y en a toujours qui dictent leurs traductions, mais de moins en moins. La formation des traducteurs reflète l'évolution de la pratique dans le métier. Dès la première année des études, l'étudiant apprend de se servir des aides informatiques, et si on privait des ordinateurs les étudiants en fin de leur formation, on risquerait de provoquer une révolution estudiantine. L'arrivée des logiciels de dictée performants, nous obligera-t-elle de faire marche arrière et de revenir à la dictée au niveau de la formation?

Pour le moment, les systèmes de dictée connaissent encore quelques limites: ils sont toujours plus performants quand ils peuvent prendre l'habitude de la voix de chaque locuteur et il est toujours nécessaire pour cette même raison de prévoir un certain temps avant de pouvoir escompter les meilleurs résultats. Mais les versions les plus récentes exigent un temps d'entraînement relativement réduit, et même sans aucun entraînement produisent des résultats nettement meilleurs que dans le passé. Pour le moment, alors, nous sommes en attente d'une réponse : mais même si le traducteur revenait à la dictée pour certaines de ses traductions, il y a des applications informatiques dont il ne peut pas se passer.

Les mémoires de traduction

La plus importante de ces applications était également une technologie en voie de développement il y a dix ans. Il s'agit des logiciels qui savent tirer bénéfice des archives de traductions déjà faites, connus sous le nom de *mémoires de traduction*. Les textes et leurs traductions qui alimentent les mémoires sont préparés par un processus d'alignement qui essaie d'établir une correspondance entre chaque phrase dans le texte source et une phrase dans la traduction. Bien que très simple à première vue, la tâche de mise en correspondance est loin d'être facile. Pour des raisons propres à la transmission du message du texte source dans la langue cible, le traducteur peut utiliser deux phrases au lieu d'une ou une phrase au lieu de deux, changer l'ordre des phrases dans le texte ou même, dans certains cas extrêmes, ajouter des informations à un texte. Si l'alignement est fait pendant que le traducteur accomplit sa tâche, les différences de structure entre le texte et sa traduction ne posent pas trop de problèmes. Mais si on veut profiter de l'existence de traductions faites dans le passé, par exemple au sein d'un service de traduction dans une grande entreprise ou dans les organisations internationales, il faut faire l'alignement en tant que procédé indépendant. Par la suite, s'il est important d'éliminer les erreurs d'alignement dues aux différences de structure, il faut contrôler manuellement les résultats de l'alignement et y porter des corrections le cas échéant. Le travail est long et ennuyeux, et implique un investissement considérable de ressources. (C'est pour cette raison que, quand les circonstances le permettent, on décide parfois d'accepter un certain taux d'erreur plutôt que de corriger scrupuleusement.)

Dès qu'une mémoire devient disponible, le logiciel compare un nouveau texte à traduire phrase par phrase avec les textes sources stockés dans la mémoire. Quand il trouve une phrase qui correspond plus ou moins exactement à une phrase déjà traduite, il recherche la traduction déjà faite et l'offre au traducteur comme traduction possible pour la phrase dans le nouveau texte. Ce processus n'a rien à faire avec la traduction automatique: le logiciel récupère des bribes de traduction faites par des traducteurs humains, et un traducteur humain décide de l'utilisation de la traduction offerte.

Il va de soi que de tels systèmes trouvent leur plus grande utilité dans la traduction de textes répétitifs. Deux types de répétition se présentent dans le travail du traducteur. Premièrement, certains textes reprennent des passages d'autres textes déjà traduits. Dans ce cas, ne pas être obligé de créer une nouvelle traduction est évidemment un gain de temps, mais parfois l'avantage peut aller encore plus loin. Il y a des cas, notamment pour certains textes législatifs, où il est très important de **ne pas** créer une nouvelle traduction: la cohérence de l'ensemble exige que le même passage soit traduit de la même façon partout – et ceci

même si le traducteur du nouveau texte n'est pas d'accord avec la traduction existante. L'autre type de répétition est caractéristique de la documentation technique: prenons le cas des logiciels comme exemple. Les logiciels évoluent très vite; les versions se succèdent à un rythme parfois presque hallucinant. Pour chaque nouvelle version qui apparaît sur le marché, il faut une documentation technique substantielle, non seulement de manuels, mais des aides en ligne, des foires à questions et ainsi de suite. Mais entre une version et son successeur, il y a des parties importantes de la documentation qui ne changent pas. Si on peut identifier ces parties et re-utiliser directement la traduction existante, on peut réduire considérablement le temps nécessaire à la production de la documentation et donc le temps nécessaire au lancement du produit sur le marché, un facteur critique pour les constructeurs et revendeurs de logiciels. Dans ces deux cas, l'utilisation d'une mémoire de traduction offre des avantages inestimables.

Les premiers systèmes incluant la possibilité de consulter une mémoire de traduction sont apparus sur le marché en 1994. Parce qu'ils répondaient à un besoin réel dans la vie de beaucoup de traducteurs, ils se sont répandus rapidement. La formation des traducteurs a suivi; l'école à Genève a incorporé les mémoires de traduction dans les cours post-grade en 1996, et à partir de 1998 chaque étudiant a eu la possibilité de se familiariser avec ces outils déjà au niveau de la licence.

Evidemment, il y a un lien entre les systèmes de mémoire de traduction et les systèmes basés sur le calcul de probabilités dans le sens que les deux profitent des nouvelles possibilités de stocker et de traiter de grands ensembles de textes. Mais il y a un autre lien encore: certains des systèmes de mémoires de traduction permettent ce qu'on appelle une *recherche floue*, c'est-à-dire qu'au lieu de chercher une correspondance exacte entre deux mots ou deux phrases, on peut aussi rechercher des éléments plus ou moins semblables: la technique utilisée ressemble au calcul de probabilités. Ainsi, en cherchant *Le conseil décide de poursuivre dans le même chemin* on trouverait *La Commission décida de poursuivre dans le même chemin*.

La plupart des systèmes de mémoires de traduction intègrent un gestionnaire de la terminologie et certains font une recherche automatique dans la base de terminologie en même temps qu'ils recherchent des correspondances dans la mémoire de traduction. Mis à part la possibilité de demander une recherche floue, la technologie de base de la plupart des gestionnaires de terminologie n'est pas vraiment nouvelle: c'est plutôt la plus grande puissance de calcul et les développements au niveau de la capacité de la mémoire des ordinateurs qui jouent un rôle décisif dans le développement de ces outils. Néanmoins, on peut voir de nouvelles technologies au niveau de l'interface utilisateur dans le cas de certains gestionnaires

qui permettent l'intégration des images ou des sons dans les fiches terminologiques, profitant ainsi des nouvelles possibilités multi-media développées surtout pour la Toile.

Les concordanciers

L'inconvénient le plus important des mémoires de traduction est indiscutablement le grand travail nécessaire si on veut profiter des anciennes traductions par l'alignement des documents archivés. A ceci s'ajoutent les frustrations de la recherche d'équivalences. Même avec un système qui permet une recherche floue, il arrive assez souvent que le système ne retrouve pas des éléments que l'être humain estime qu'il devrait trouver. L'exemple classique concerne le cas où une phrase dans le texte contient une phrase plus courte contenue dans la mémoire ou vice-versa. Disons que j'ai dans le texte:

Pour l'année prochaine, la rentrée est fixée pour le 20 octobre, et les étudiants sont priés d'assister aux premiers cours le 21.

Et dans la mémoire:

Pour l'année prochaine, la rentrée est fixée pour le 19 octobre. Les étudiants sont priés d'assister aux premiers cours le 20.

Une recherche d'équivalences ne trouverait ni la première phrase de la mémoire ni la deuxième, et le problème n'est pas la différence dans les dates, comme on peut confirmer en variant la phrase du départ. Le problème vient tout simplement du fait que chaque phrase est comparée séparément, et la comparaison travaille au niveau de la forme superficielle des phrases à comparer sans pouvoir aller plus loin. Donc, en comparant la phrase du nouveau texte à la première phrase dans la mémoire, on constate qu'il n'y a que la moitié de la phrase qui correspond – un taux de ressemblance qui ne suffit pas pour que le logiciel considère qu'une vraie ressemblance valable a été trouvée. Il en va de même pour une comparaison avec la deuxième phrase de la mémoire. Il n'est pas surprenant que les traducteurs trouvent cette faiblesse frustrante: ils sont convaincus parfois d'avoir traduit toute une phrase difficile auparavant, mais le logiciel ne trouve rien.

Afin de combler cette lacune, certains systèmes permettent une recherche par sous-phrase. Le traducteur sélectionne la suite de mots qui l'intéresse, et le logiciel fait une recherche dans toute la mémoire pour les occurrences de cette suite de mots. Comme il s'agit d'un élément plus petit qu'une phrase entière, le logiciel ne peut pas identifier la suite de mots qui correspond au niveau de la traduction. Il affiche alors toutes les phrases entières qui contiennent la suite de mots recherchée, et toutes les phrases entières de la traduction qui sont censées correspondre. Le traducteur peut alors choisir parmi les phrases identifiées celle qui

contient la traduction qu'il veut et avec une opération de copier-coller insérer la sous-phrase choisie dans sa traduction.

Récemment, des systèmes qui vont plus loin dans cet ordre d'idées sont apparus sur le marché. Ces systèmes ne créent pas une mémoire de traduction, mais travaillent plutôt directement à partir d'un archive de textes. Le logiciel fait un alignement automatique, mais l'utilisateur ne corrige pas l'alignement avant de commencer la recherche des suites de mots qui l'intéressent. Pour certains de ces systèmes, il peut quand même corriger les fautes d'alignement au fur et à mesure qu'il les trouve.

Il est intéressant de noter, de nouveau, que l'idée d'une recherche en plein texte est loin d'être nouvelle: une des toutes premières applications de l'informatique dans le traitement automatique de la langue était l'élaboration de concordanciers automatiques, beaucoup utilisés dans les études de style, de l'utilisation de la langue et d'attribution d'auteur. La nouveauté dans les systèmes actuels est l'utilisation de bi-textes et la tentative d'identification du paragraphe ou de la phrase dans la traduction qui correspond à un passage dans le texte source.

Il est encore un peu tôt pour prévoir l'avenir de ces systèmes. Ils ont certes l'avantage d'éviter la corvée d'un alignement fastidieux et coûteux, mais au coût de devoir demander explicitement une recherche chaque fois et d'être obligé de faire du copier-coller pour transférer les résultats de la recherche à la traduction. Peut être la réponse est "tout va dépendre du type de textes à traiter", comme c'est souvent le cas avec les outils d'aide à la traduction. Celui qui travaille avec beaucoup de textes répétitifs va peut être préférer une mémoire de traduction, celui qui traduit des textes moins répétitifs un concordancier.

La Toile et la gestion des informations.

Une dernière technologie en voie de développement il y a dix ans qui s'est énormément développée depuis et qui influence profondément la vie des traducteurs est la Toile elle-même. Nous avons déjà remarqué que la Toile de l'époque était surtout une affaire de scientifiques, qui l'avait inventée dans un but de partage des informations techniques. Ces dernières années ont vu la mise à disposition des informations de toutes sortes sur la Toile, couplé avec le développement de moteurs de recherche efficaces et conviviaux qui permettent à tout un chacun de chercher les informations qui l'intéressent. Deux autres facteurs ont joué un rôle critique dans l'extension de la Toile vers le grand public. D'abord, de nouvelles techniques permettent une présentation des informations beaucoup plus agréable et facilitent leur assimilation. Ensuite, la nature même des informations disponibles a changé: au début les

informations étaient surtout celles qui passaient par la langue écrite, maintenant l'information est transmise par les graphiques, les images, l'animation, les sons, la vidéo et la musique. Ajoutons encore que l'utilisation de la Toile a changé: au lieu d'utiliser la Toile seulement comme un moyen de satisfaire nos besoins en information nous commençons à y transférer d'autres activités. Nous utilisons l'Internet pour faire nos opérations bancaires, pour faire des achats, pour écouter de la musique, pour nous amuser en jouant, et pour bien d'autre chose.

Tout ceci est autant vrai pour un traducteur que pour tout le monde. Mais c'est peut être la recherche des informations qui joue toujours le plus grand rôle au niveau de sa vie professionnelle. Imaginons, par exemple, le traducteur à la recherche d'un bon équivalent pour un terme qu'il vient de rencontrer dans un texte. Avec l'aide de l'Internet, il peut consulter des journaux en ligne, regarder les publications spécialistes, chercher dans des dictionnaires et bases de terminologie à distance et, dans le cas où il ne trouverait rien, discuter le problème avec des collègues géographiquement éloignés.

Ces deux derniers paragraphes donnent un peu l'impression d'une utopie où la solution à tous les problèmes de renseignement est presque directement sous la main, ce qui est loin d'être reflété dans l'expérience de tous les jours. La difficulté actuelle est l'énorme quantité d'information disponible et sa gestion. Si nous avons tous trouvé la solution à un problème en un clin d'œil, nous avons tous également fait l'expérience d'une recherche qui résulte dans l'offre de trois mille documents. On retournera à ce sujet dans le dernier chapitre de cet article.

Les technologies en voie de développement aujourd'hui.

Nous avons commencé par un paradoxe: en regardant les dix dernières années, on constate qu'il n'y a pas eu de changement révolutionnaire dans les bases théoriques sous-jacentes aux aides à la traduction, mais qu'en même temps, la vie quotidienne des traducteurs a profondément changé. En 1993, le traducteur qui se servait beaucoup de l'informatique faisait image de pionnier et futuriste. En 2003, le traducteur qui n'utilise pas l'informatique est en voie de disparition. Il est vrai que certaines technologies qui étaient très nouvelles en 1993 sont devenues mûres entre-temps, mais la vraie résolution du paradoxe réside, je crois, dans le développement du matériel informatique, dans la production quasi universelle de documents sous forme électronique et dans le développement de l'Internet.

Ces mêmes facteurs jouent un rôle clé dans les technologies en voie de développement d'aujourd'hui. Tout d'abord, je suis convaincue que nous ne sommes qu'au début de la révolution provoquée par la création de la Toile. Nous avons remarqué plus haut que le gros problème actuel est la maîtrise des informations disponibles. Les moteurs de recherche ont

déjà fait des grands progrès, mais ils restent essentiellement basés sur une recherche des mots clés. L'utilisateur donne quelques mots, et le moteur cherche tous les documents qui contiennent ces mots. Bien sûr, certains moteurs permettent aux utilisateurs de spécifier que les mots doivent se suivre dans l'ordre donné ou doivent se présenter exactement sous la forme spécifiée. D'autres permettent une combinaison des mots clés avec des opérateurs logiques; on peut demander qu'un document contienne ce mot-ci mais pas ce mot-là par exemple. Il y a même des moteurs qui se reposent sur une classification de thèmes (normalement construite manuellement) qui sert à guider la recherche et en même temps à limiter le nombre de résultats offerts à l'utilisateur. Mais on peut aller beaucoup plus loin et imaginer des moteurs qui collaborent avec l'utilisateur dans un effort de définir ses intentions et ses besoins et qui font, en effet, une recherche intelligente des informations.

Pour le moment, la recherche multilingue aussi est assez limitée. L'utilisateur peut spécifier la langue des informations qu'il veut et il peut demander une traduction automatique de certains des documents trouvés. De nouveau, on pourrait envisager d'aller plus loin, en imaginant qu'on pourrait extraire d'une demande d'information exprimée dans une langue spécifique l'essentiel de son sens afin de chercher les informations voulues même si elles sont contenues dans un document exprimé dans une autre langue. Un simple exemple serait une requête en français pour les heures d'ouverture des musées en Espagne, où la réponse est trouvée sur un site en langue espagnole.

Mais de nouveau toutes ces idées ne sont pas fondamentalement nouvelles: elles sont en quelque sorte la suite logique des moteurs de recherche que nous connaissons déjà. La grande évolution viendra, je crois, avec des systèmes qui essaient d'exploiter la masse même des informations disponibles et de réaliser des exploits que seule l'informatique peut faire. Prenons un exemple très simple. Il n'est pas possible pour un être humain de lire tous les journaux produits en France chaque jour et de se rappeler de leur contenu. S'il pouvait se créer une maîtrise de toutes les informations contenues dans tous les journaux, il pourrait probablement découvrir des liens entre certaines informations qui ne sont pas évidents ni même trouvables sans cette maîtrise. Ce que l'homme ne peut pas faire, l'informatique peut: avec les progrès en puissance de calcul et en mémoire, il devient tout à fait envisageable qu'un logiciel digère tous les journaux chaque jour et mémorise leur contenu. Le défi est dans la création de logiciels qui savent exploiter les informations mémorisées. Il y a actuellement beaucoup de recherche dans ce sens, dans les domaines comme l'acquisition des connaissances (*knowledge discovery*), l'exploration du texte (*text mining*) et l'exploration des données (*data mining*). L'intérêt de tels systèmes est évident et va du responsable du

supermarché qui décide de ranger deux produits ensemble sur la base d'une découverte dans les données que celui qui achète un des deux achète souvent l'autre, jusqu'au politicien qui décide de proposer la construction d'une piscine dans une banlieue parce que les données suggèrent une association entre la diminution de la violence et l'accès à une piscine. Les logiciels qui existent déjà sont surtout exploités aux fins commerciales et militaires. Mais comme le développement de la Toile nous l'a montré, l'affaire des spécialistes aujourd'hui peut changer la vie quotidienne de tout le monde demain.

Les avances dans les technologies de communication également profitent à tout le monde, et non seulement aux traducteurs. Ici nous voyons des nouveautés pratiquement toutes les semaines. Ces avances ont quand même déjà provoqué des changements profonds dans l'organisation du travail de certains traducteurs. Il y a quelque temps, les entreprises et les grandes organisations travaillaient de préférence avec leurs propres équipes de traducteurs sur place. Dès que la communication par voie électronique est devenue rapide, fiable et accessible à tout le monde, certaines agences de traduction ont vu le jour où presque tout le travail de traduction est fait par des traducteurs indépendants, souvent géographiquement éloignés du siège de l'entreprise. Ceci est particulièrement vrai pour les entreprises spécialisées en localisation, c'est-à-dire le processus par lequel un produit (la documentation technique par exemple ou un site sur Internet) est transformé afin de l'adapter à une culture spécifique. La transformation concerne non seulement la langue, mais également tous les autres éléments spécifiques à une culture, les couleurs, par exemple, ou la façon d'écrire les chiffres ou l'utilisation des images. Il est très important que celui qui fait l'adaptation reste dans le bain de la culture visée, d'où l'intérêt pour l'entreprise de garder un réseau de traducteurs, qui, eux, travaillent chez eux par courrier électronique et par Internet ne voyant peut être jamais le siège de l'entreprise. (voir aussi le site du LISA, l'association pour la standardisation dans l'industrie de localisation). On peut prévoir une expansion de ce type d'organisation du travail.

D'autres évolutions vont concerner les traducteurs qui utilisent déjà les outils d'aide à la traduction. Des recherches sont en cours, par exemple, sur les algorithmes d'alignement de bi-textes. Nous avons vu ci-dessus qu'un des inconvénients des aligneurs actuels est qu'ils travaillent au niveau de la phrase complète, ce qui empêche la découverte d'équivalents au niveau de sous-phrases ou au niveau des mots individuels. Toute amélioration va profiter directement à celui qui utilise une mémoire de traduction ou même un concordancier.

Un autre domaine où la possibilité de faire un alignement plus fin pourrait changer la qualité des résultats obtenus et ainsi influencer la vie quotidienne des traducteurs est l'extraction de la terminologie. Avec les avances continues de la technologie dans tous les

domaines il y a de nouveaux termes techniques inventés tous les jours. Le traducteur confronté par un nouveau terme est parfois obligé de passer un temps considérable dans la recherche avant qu'il trouve un équivalent. Étant donné que le terme et sa traduction existent peut-être déjà dans la masse de nouvelle documentation qui apparaît également tous les jours, il serait souhaitable de faciliter la tâche du traducteur en y faisant l'extraction automatique de tous les nouveaux termes et de leurs équivalents. Certains produits d'extraction sont déjà sur le marché, mais les résultats ne sont pas très satisfaisants. Le problème est un problème de *bruit*, à deux niveaux. Au niveau du texte source, le logiciel d'extraction extrait des suites de mots qui ne sont pas des termes. Au niveau du lien entre le texte en langue source et le texte en langue cible, le logiciel ne réussit pas à trouver le mot ou la suite de mots qui est réellement la traduction du terme dans le texte source. Une amélioration dans les algorithmes d'alignement apporterait au moins partiellement une solution au deuxième problème. Le premier est peut-être plus difficile; avec quelques raffinements au niveau linguistique, il est possible de bloquer une proportion des mauvais résultats. Mais pour arriver à un résultat parfait, il faudrait pouvoir définir explicitement des critères pour ce qui constitue un terme, et là on se heurte de nouveau à la barrière sémantique.

On ne peut pas conclure un tel tour d'horizons sans parler de l'avenir de la traduction automatique. En regardant le passé, nous avons distingué deux types de systèmes, les systèmes basés sur les jeux de règles linguistiques et les systèmes basés sur un calcul de probabilités. À un certain moment il y avait même un grand débat qui confrontait les deux types de systèmes. La tendance actuelle est de créer des systèmes hybrides, qui utilisent des règles et des informations statistiques (Charniak et al, 2003).

Pourtant un nouveau défi s'est déclaré assez récemment où les systèmes empiriques semblent avoir un net avantage sur les systèmes basés sur les règles linguistiques. Il s'agit de la création rapide de nouveaux systèmes de traduction automatique. L'élaboration de jeux de règles linguistiques exige un travail long et fastidieux. Par contre, si un corpus de textes parallèles est disponible, la création rapide d'un nouveau système basé sur les théories statistiques semble faisable (Foster et al, 2003). Il y a même eu des recherches récentes sur la création de systèmes de traduction automatique pour des langues inconnues à celui qui construit le système (Oard et Och, 2003).

Finalement, des systèmes qui intègrent la reconnaissance et la génération de la parole avec la traduction automatique commencent à apparaître (Raynor et Bouillon, 2002, 2003). Il ne faut pas confondre ces systèmes avec des systèmes qui font l'interprétation. L'interprétation est un processus où un discours parlé est directement transformé dans un discours écrit dans

une autre langue: l'interprète travaille en temps réel sans avoir recours à une version par écrit du discours ni le temps de réflexion dont un traducteur peut bénéficier. Avec un système de traduction de la langue parlée, il y a une première étape pendant laquelle l'énoncé est traité par un logiciel de reconnaissance de la parole et transformé en version textuelle. Le module de traduction prend cette version textuelle comme point de départ, et traduit vers une représentation textuelle dans la langue cible, qui sert ensuite comme base pour la génération de la version parlée de la traduction. Les systèmes de traduction de la langue parlée qui existent déjà travaillent tous dans un domaine limité, tel que la consultation d'un médecin. Mais même dans ces limites, l'utilité potentielle de ces systèmes est indéniable.

Conclusion.

Il serait impossible de faire un résumé de tout ce qui a été discuté dans cet article, bien que la discussion soit restée à un niveau assez superficielle et, j'espère, pas trop technique. La seule conclusion possible alors est que les changements constatés sur les dix derniers ans présagent des changements encore plus grands et encore plus rapides. L'évolution est claire: la révolution ne vient que de commencer.

ISSCO/TIM/ETI

Uni-Mail

40 blvd du Pont d'Arve

CH-1211 Genève 4

Tél: +41 22 379 87 55

Margaret.King@issco.unige.ch

Bibliographie.

Abrams, M, ed. (1998) *World Wide Web – Beyond the Basics*, Prentice Hall.

Church, K.W. et Hovy, E.H. (1993) Good applications for crummy MT. *Machine Translation* 8, 239 – 258.

Ananiadou, S. (1987). A Brief Survey of Some Current Operational Systems. King, M. (ed) *Machine Translation Today*, 171 – 191, Edinburgh, Edinburgh University Press

Charniak, E., Knight, K., Yamada, K. Syntax based language models for statistical machine translation. Actes MT Summit IX, 2003. New Orleans, International Association for Machine Translation.

Dien, D., Hoang, K., Hovy, E., (2003). Actes MT Summit IX, 2003. New Orleans, International Association for Machine Translation.

Foster, G., Gandrabur, S., Langlais, P., Plamondon, P., Russell, G., Simard, M. (2003). Statistical machine Translation: Rapid Development with Limited Resources. Actes MT Summit IX, 2003. New Orleans, International Association for Machine Translation.

Hutchins, W.J. (2000). *Early Years in Machine Translation*. Amsterdam, John Benjamins.

Isabelle, P. (1987). Machine Translation at the TAUM Group. King, M. (ed) *Machine Translation Today*, 247 – 277, Edinburgh, Edinburgh University Press.

Lebert, M., (1999) *Multilingualism on the web*.

A l'adresse: <http://www.cefrio.qc.ca/projects/Documents/mutlieng2.htm#21>

LISA: <http://www.lisa.org>

Nirenburg, S., Somers, H. et Wilks, Y. (2003). *Readings in Machine Translation*. Boston, MIT Press, a Bradford Book.

Oard, D, Och, F. (2003). Rapid Response Machine Translation for Unexpected Languages. Actes MT Summit IX, 2003. New Orleans, International Association for Machine Translation.

Raynor, M. et Bouillon, P. (2002). A phrasebook style speech medical translator. *ACL-02 Companion Volume to the Proceedings of the Conference*, Section 2, 114-115. Philadelphia. ACL.

Raynor, M., Bouillon, P., Van Dalsem V., Isahara, H., Kanzaki, K et Hockey, B.A. (2003). A limited-domain English to Japanese medical speech translator built using Regulus 2. *ACL-03 Companion Volume to the Proceedings of the Conference*.

Raynor, M., Carter, D., Bouillon, P., Digikalis, V et Wirén, M. (eds) (2000). *The Spoken Language Translator*. Cambridge, U.K., Cambridge University Press.

Sauron, V, (2002). Tearing out the terms. Actes, ASLIB. London, ASLIB.

Wayne, C., 2002. Transforming Fantasy, *DARPA Tech Symposium 2002*.

A l'adresse : <http://www.darpa.mil/darpatech2002/presentations>